THE UNIVERSITY OF WARWICK

University of Warwick institutional repository: http://go.warwick.ac.uk/wrap

A Thesis Submitted for the Degree of PhD at the University of Warwick

http://go.warwick.ac.uk/wrap/58425

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

The Epiphenomenal Mind

by

Simon Buttars

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Sociology

University of Warwick, Department of Sociology

September 2003

Contents

***Part I*: Consciousness and Causation**

Figures

Acknowledgements

This thesis was written with the aid of an Economic and Social Research Council training award. Grateful thanks are due to the ESRC. Thanks are also due to Professor Margaret Archer and Professor Roger Trigg for their objective approach when supervising what is certainly an unorthodox thesis.

Abstract

The Epiphenomenal Mind is both a deflationary attack on the powers of the human mind and a defence of human subjectivity. It is deflationary because in the thesis I argue that consciousness is an epiphenomenal consequence of events in the brain. It is a defence of human subjectivity because I argue that the mind is *sui generis* real, irreducible, and largely an endogenous product (i.e. not dependent on society or its resources).

Part I is devoted to arguing that the conscious mind is epiphenomenal. Arguing from, the irreducibility of mental states, the causal closure of the physical domain, and the principle of causal explanatory exclusion, I seek to demonstrate that all theories of mental causation necessarily violate one or more of these premises. Contemporary approaches to mental causation come under two broad categories, those that argue that mental events are supervenient on physical events (such as Davidson, Kim and Horgan) and those (like Haskar) who argue that the mind is an emergent property of the brain. Supervenience based theories, I argue, end up reducing mental states in their search for a theory of mental causation and emergence based theories end up violating the principle of the causal closure of the physical.

In part II, I explore some of the consequences of epiphenomenalism for social theory. This exploration comes in the context of a defence of human subjectivity against (i.) those sociological imperialists who view the mind and self as a 'gift of society', and (ii.) social situationalists who have abandoned the concept of action and an interest in 'what's in the head' of the actor, in favour of a concept of social action which views behaviour as action only to the extent that it is socially meaningful. The conclusion is that the social sciences should return to an interpretative style (Weberian) methodology.

Abbreviations

AM anomalous monism

CCP causal closure of the physical

FFP the folks' folk psychology

MMKE mean molecular kinetic energy

OMR orthodox mental realism

P1 premise 1: The irreducibility of mental states

P2 premise 2: The causal closure of the physical

P3 premise 3: The principle of causal explanatory exclusion

PE phenomenal experience

PFP the philosophers' folk psychology

RP readiness potential

SS strong supervenience

Part I: Consciousness and Causation

Chapter 1

Making Mind Matter Less

My itching isn't responsible for my scratching, my wanting a glass of wine did not cause me to open a bottle, and my belief that it is raining had nothing to do with my decision not to go riding today. Practically everything I believe about anything is false, but it's not the end of the world. It's not even the end of predicting and explaining people's behaviour.[1] Rather, it's the start of understanding ourselves and our place in the world better, and it's an acknowledgement we must make if we want to be able to explain the real causes of people's actions.

As the title of this thesis suggests, I believe that conscious experience is an epiphenomenal consequence of events in the brain. That is to say, our conscious mental states (the sensation of pain, the anticipation that awaits the arrival of an old friend or eating one's favourite meal, the addict's cravings, and so on) have no causal powers. They are all the result of events in the brain, but the *experiences* themselves do nothing and it would make no difference to the way that we behave if we had no conscious life at all!

Before exploring this position in detail, and the considerations that led to my adopting epiphenomenalism, it might be helpful to outline exactly what is being denied here. Being conscious, or so the orthodox story goes, differentiates humans, and certain higher order animals, from inanimate objects and simple forms of life (plants, single celled organisms, insects and the like). At some point during the evolutionary history of life a magic level of complexity was reached in the brains of some species and consciousness 'popped out' (we'll leave worrying about how this should be phrased until later). It may have just been a glimmer at first, some perception or sensation, certainly nothing like the fully-fledged self-consciousness we all experience. Whatever its extent though, it conferred some evolutionary advantage on the organisms in which it was instantiated. *Feeling* pain made it more likely that organisms would withdraw from and avoid harmful stimuli. *Remembering* the location of danger or food conferred an evolutionary advantage on the individual/group/species (depending on one's preferred theory regarding the unit of selection) that possessed the knowledge. And the ability to *reason* and *imagine* allowed conscious organisms to predict the consequences of their actions, allowing theories 'to die in their stead'. Gradually consciousness grew in both intensity and complexity until humans ultimately developed fully-fledged self-consciousness. We developed the capacity to think reflexively about ourselves, to distinguish ourselves from others and to realise that other humans are also self-conscious beings. In all these cases, so the story goes, the *feeling* of consciousness is a necessary condition for the behaviour that it *causes*. Thus, I don't open a bottle of wine because some neurons in my brain fire thus and so, I open a bottle of wine because I have a *conscious desire* for a glass of wine. I imagine the taste, smell, and feelings of pleasant intoxication that will result from its consumption. The relationship between neural events and conscious experience (or, to avoid any ambiguities, what we will term phenomenal experience) is, of course, highly problematic and

[1] Those of you familiar with the literature on mental causation will recognise the title and opening lines of this chapter as an alternative version of the last paragraph of Fodor's Making Mind Matter More.

goes to the heart of the mind-body problem. At the outset, though, it is crucial to stress that both the lay and philosophical conceptions of mental causation hold that the phenomenal experiences (the 'what it is like' to desire, taste, see, touch, etc.) of mental states are a necessary condition for their causal efficacy. In laymen's terms, it does not matter a jot what my brain is doing, if I had not imagined the taste and smell of a bottle of Australian Shiraz and thought how well it would go with my dinner, I would not have opened a bottle. Henceforth we shall term this position 'orthodox mental realism.' I add the prefix 'orthodox' here because, despite advocating epiphenomenalism, the position I will defend in this thesis is still one of mental realism. That is to say, contrary to eliminative materialists (whose position we shall encounter later), I will argue that we must accept the world of phenomenal experience just as we find it and not attempt to reason it out of existence. It is not the existence of phenomenal experience which epiphenomenalism denies, only its causal powers.

Over the course of this thesis I will develop an account of epiphenomenalism in which one's conscious experience is analogous to one's shadow. The motions of a shadow are a consequence of the movements of the body, just as the phenomenal experience of mental states is a consequence of neural events. The shapes of shadows are determined by the shapes of the objects creating them, just as the structure, content, and phenomenology of conscious experiences are determined by the neural events that cause them. Moreover, just as it is a physically necessary condition that an object casts a shadow (under the appropriate circumstances), it is a physically necessary condition that certain neural events instantiate the phenomenal experiences they do. That is to say, it is a brute fact of nature that certain neural events cause certain phenomenal experiences.

Most philosophers base their understanding of the mind-body problem on their intuition that the mind causally interacts with the body: that what they think and feel has a causal influence on their body and, thereafter, on the world. They are convinced that it is their pain that causes them to wince, writhe, scream, cry out, etc., and not the neurophysiological events going on in their brains. Feelings of pain, lust, anger, hunger, and so on, seem to grab hold of us and demand that we act. The idea that these feelings don't do anything, that it would make no difference to the way we behave if we never experienced them, seems absurd. Although almost everybody thinks that mental states interact causally with the body, discovering the nature of this interaction has proved to be an intractable problem.

Neither the dualist nor the materialist camp has produced anything approaching a solution. Dualists can't explain how mind affects matter (or to the materialists' consternation believe they don't need to) and materialists constantly face the embarrassment of qualia.[2] Nevertheless, protagonists of both camps agree on the problem (explaining how mind affects matter) and the phenomenon under investigation (the material world as described by physics and the phenomenal world of experience). The battle, it seems, will be won or lost over qualia. The materialists' strategy for securing their efficacy has been to reduce or identify qualia with matter. The dualist camp, in contrast, has adopted the diametrically opposed position, maintaining that the only way to save their causal efficacy is to insist that qualia are *sui generis* real with irreducible causal powers.

Epiphenomenalism, it is almost universally agreed, has an unhappy alliance with both camps. On the one hand it refuses to agree with materialists that qualia are reducible to or identical with physical states, while at the same time refusing to acknowledge their causal powers. Consequently, qualia are left dangling – portrayed as *sui generis* real, but entirely inefficacious. By claiming that phenomenal experiences are inefficacious, epiphenomenalists deny themselves the standard Darwinian explanation for their existence and lay themselves open to the three standard arguments against epiphenomenalism. Viz. that it cannot explain: (i.) the correspondence between the function and feel of mental states, (ii.) how a putatively biologically expensive but inefficacious property could evolve, and (iii.) why one should take seriously an argument that, since it claims to be merely the effect of physical processes, cannot appeal to reason for its justification. The first two of these problems will be discussed in chapter 5, where I will also develop and deal with some further challenges to my position. The third objection will be discussed when we consider rationality in chapter 8.

Epiphenomenalism is an extremely difficult position to argue for directly.[3] There are a handful of neurologists and cognitive scientists who attempt to use experimental evidence to argue that mental states are epiphenomenal. This usually takes the form of showing that certain neural processes believed to be the cause of behaviour occur prior to our conscious decision to act. Research has focussed on the role of focal-attentive processing of

[2] Qualia refer to what I have been calling the phenomenal experience of mental states.

[3] The only recent advocate of epiphenomenalism (of which I am aware) is Keith Campbell (1970). Though rather unsophisticated by modern standards, Campbell argues from the irreducibility of phenomenal properties and the completeness of physics for a 'new Epiphenomenalism'. The new Epiphenomenalism is a form of double aspect theory which accepts that the mind and body causally interact (because the mind *is* the brain) but claims that phenomenal properties are non-physical effects of brain processes and entirely inefficacious.

complex stimuli, learning, decision making, and activities that require some degree of planning. Some researchers have argued that consciousness occurs as a consequence of these processes rather than causally entering into them. The evidence is far from decisive though and opponents of epiphenomenalism have used the same material to argue for mental causation.[4] Since the empirical route is closed, and phenomenology is on the side of the orthodox mental realists, our only option is to eliminate the alternatives and demonstrate that epiphenomenalism is the only position consistent with a (broadly) materialist perspective. Our main goal in part I will be to evaluate the full range of positions open to materialist theorists of mental causation. I will argue that, although there is still a lot of work to be done putting flesh on the bones, all the metaphysically possible positions on mental causation have been identified and none are consistent with the fundamental principles of materialism (outlined below). This conclusion presents us with a stark choice: either abandon our belief in mental causation and adopt epiphenomenalism or drop one or more of our materialist principles. This is uncharted territory for the philosophy of mind and we have few principles with which to guide our choice. Short of (an unlikely) scientific refutation, we have little reason to abandon these three principles save for their contradicting our deeply held intuitions regarding the causal efficacy of mental states. Though neither option is appealing, I have opted for the former.

[4] For a summary of this research see Velmans (1991) paper and subsequent peer commentary.

(P1) The irreducibility of phenomenal states

In what follows I will only mention the two anti-reductionist arguments that I find most convincing: the argument from property dualism, and what I shall term the causal mereological argument. Since reductionist accounts of mental causation are now generally regarded as having failed, we need only briefly consider the anti-reductionist arguments here. Non-reductive physicalism, which now dominates philosophy of mind, will be the subject of part I.

The most familiar anti-reductionist argument exploits our intuitions regarding the incommensurability of physical and phenomenal properties. The argument, which has been most forcefully expounded by Thomas Nagel (1995), Frank Jackson (1982), and John Searle (1992), runs as follows: even if we know everything there is to know about the neurophysiology of a conscious state (be it the neurophysiology of a bat's echolocation or the experience of seeing red), we still do not know what it is like to experience the corresponding conscious state. As Searle notes (1992: 116-7), this is often interpreted as an epistemological argument when it should be read as ontological. The force of the argument derives from the fact that if we try to make an ontological reduction of qualia to their neurophysiological 'cause' something is always left out. With most ontological reductions, according to Searle, we first achieve a causal reduction (i.e. we discover that the causal powers of the reduced property are explained by the causal powers of the reducing property), before redefining the reduced property in terms of the reducing. The standard examples of this type of reduction include heat to mean molecular kinetic energy, light to electromagnetic radiation, colours to photon emissions of particular frequencies, and solidity to molecular movements within a lattice structure. In such cases we begin with a definition couched in terms of its subjective experience (the appearance) before redefining the phenomenon after scientific discoveries (the reality). In the case of conscious experience, however, we cannot make this appearance-reality distinction 'because consciousness consists in the appearances themselves. *Where appearance is concerned we cannot make the appearance-reality distinction because the appearance is the reality*' (Searle 1992: 122). To explicate, we may say that to experience heat is to experience mean molecular kinetic energy (MMKE), but the sensation of heat is *not* 'nothing but' a neurophysiological response to MMKE. The experience of, for example, a roaring fire on a winter's evening is constituted by phenomenal properties such as warmth, pleasure, contentment etc. that are inevitably missed out in a neurophysiological description of the cause of the experience.

What I have termed the causal mereological[5] argument runs on a slightly different track. This argument exploits our Cartesian intuitions regarding *res cogitans* and *res extensa*[6] in the context of mental causation. Let us suppose that token physicalism, which is a prerequisite for reductionist theories of mental causation, is true, and grant, for the sake of this argument, that mental states are causally efficacious. Thus, we are assuming that a token mental state X is identical to a physical state P_1 (we will call this 'composite' state XP_1). Now let's suppose that XP_1 causes YP_2, that is to say, token mental state X, which is identical to physical state P_1, causes mental state Y, which is identical to physical state P_2. We can now exploit the Cartesian intuition by working through the consequences of our second principle, the causal closure of the physical. As we

[5] Mereology concerns the relations between parts and wholes.

[6] Literally, 'thinking thing' and 'extended thing'.

shall see this principle states that all causation is physical causation, and arguably entails that one can provide a complete causal account that cites only the most basic physical particles, their properties and relevant laws. Emergentist theories of mental causation will be the subject of chapter 3, so we needn't worry that this scenario begs the question against emergentism. In any case emergentism denies that qualia are reducible and it is reductionist theories which concern us here. We are assuming that P_1 and P_2 are 'nothing but' an aggregate of neurons in two different configurations and that the state of each neuron in P_2 can be explained by the behaviour of the relevant neurons in P_1. Now if X is identical to P_1, and certain parts of P_1 caused P_2, it must be the case that certain parts of X can be identified with certain parts of P_1. But this is patently false, for while it may be plausible to suggest that certain mental states are composites of different perceptions, sensations, cognitions, etc., it would be absurd to claim that mental states are divisible into hundreds, thousands or perhaps millions of constituent elements and that each of those elements is identifiable with a part of P_1 (say, for example, the firing of a single neuron or the movement of a potassium ion across a cell membrane). Furthermore, even if such an identification were possible, the identity thesis would still face the problem of accounting for the unity of conscious experience *within the causal framework*. That is to say, on this model, XP_1 is conceived as an aggregate of microphysiological events together with their phenomenal correlates. According to our second principle, XP_1's causing YP_2 is only shorthand for the constituent elements of XP_1 cause the constituent elements of YP_2 to move into the specific configuration that composes YP_2. However, if this were the case then it would make the characteristic unity of consciousness as well as the idea of a self (which depends on the unity of consciousness) epiphenomenal. That is to say, it would make no causal difference if we experienced the world as a disjunction of separate percepts rather than as a unified whole. Such an impoverished concept is surely not what defenders of mental causation have in mind.

(P2) The causal closure of the physical

In its basic form (which is adopted by Kim) this principle states that everything that has a cause has a physical cause. As such it is principally designed to rule out dualist interactionism, but it is formulated in such a way that it does not exclude the possibility of purely random events. Lynne Rudder Baker, however, argues for a more precise formulation of the thesis which she expresses as follows:

> Every instantiation of a micro-physical property that has a cause at *t* has a complete micro-physical cause at *t*. (Baker 1987: 79)

Baker argues that a system is causally closed if and only if the elements of the system interact only with other elements of the system. She therefore claims that, for example, neurophysiological systems are not causally closed (which she notes is an assumption held by many in the debate on mental causation) because 'lower level' phenomena such as molecular or quantum events will causally influence neurophysiological processes that would otherwise be governed by neurophysiological laws. By her standards the only system that could count as causally closed is the microphysical, 'where "micro-physical" is a name for whatever turn out to be basic physical particles and their properties' (ibid.). Baker, I should note, has a vested interest in formulating the thesis in this manner because her project is to argue that the metaphysical assumptions underlying materialism necessarily preclude a solution to the problem of mental causation.

The argument that only the microphysical domain is causally closed is essential for this project which concludes that we should ditch the fundamental premises of materialism.

Baker's restrictive definition is easily avoided by thinking of the physical domain as a whole as the system which is causally closed rather than any particular 'level'. Such a move seems entirely justified since the causal closure of the physical is an ontological thesis that is entirely insensitive to the epistemological definition of, for example, the 'neurophysiological system'. By retaining a definition of causal closure that does not apply solely to microphysical systems we also avoid begging the question against emergentist theories of mental causation (which will be discussed later).

Nevertheless, despite taking issue with Baker's logic, a strong case can be made in favour of her formulation if one adds strong supervenience as a premise. Oddly strong supervenience is the second assumption Baker identifies as being held by materialists which she argues makes the problem of mental causation insoluble. It is strange, therefore, that she chose not to use supervenience as a premise from which to argue for her formulation of the principle of causal closure. I will make the case for Baker's formulation below when I discuss the metaphysics of causation but at this point I want to move on and outline the third premise.

(P3) The principle of causal explanatory exclusion

This principle states that there cannot be two *complete* and *independent* causal explanations which share the same explanandum (see Kim 1993b). This is a relatively straightforward principle that is derived from the requirement that causes be counterfactually necessary for their effects. Thus, as Kim phrases it: 'If

C is sufficient for a later event E, then no event occurring at the same time as C and wholly distinct from it is necessary for E' (ibid. 243). This principle does not rule out the possibility that we may have several different and correct explanations for the same event, such as a rationalising explanation and a physiological explanation. It does, however, rule out the possibility that the rationalising explanation and the physiological explanation could both be *complete* and *independent* causal explanations. Consequently, whenever we confront a situation where we have two or more causal explanations, we are forced to enquire as to the relationship between the explanations in order to ensure that the principle has not been violated.

The potential tension between the belief that mental states are causally efficacious (and hence that rationalising explanations are valid) and these three principles is the central problem confronting contemporary philosophy of mind. Those theorists that have attempted to secure the causal efficacy of mental states within an anti-reductionist framework have gravitated towards supervenience or emergence. Crudely speaking, these perspectives represent the contemporary incarnations of materialism and dualism respectively. However, the traditional versions of materialism and dualism are long dead and the contemporary inheritors of these traditions have spent the last fifty years squabbling over the spoils. The result is, frankly, something of a mess, with someone or other representing every conceivable point on what is now a continuum between the two (previously diametrically opposed) positions. Despite claiming that supervenience and emergence are the contemporary incarnations of materialism and dualism, it would be wrong to suggest that one could place these perspectives at opposing ends of the spectrum. Rather than pointing to any concrete set of theories these positions are best viewed as approaches to the mind-body problem based loosely around the issue of reductionism. Where supervenience theorists tend towards a more reductionist strategy (while assiduously avoiding the now unpopular term), emergentists stress the irreducibility of mental states. It is my contention, though, that neither approach has succeeded (or indeed can succeed) in relieving the tension. Supervenience, I shall argue, violates our first premise (the irreducibility of mental states) and emergence our second and/or third premises. It is in this sense that supervenience and emergence may be seen as the contemporary versions of materialism and dualism, since in violating these premises their theories collapse back into the very traditions they sought to supersede.

Following our discussion of supervenience and emergence we shall turn our attention towards the contemporary debate over connectionism and its implications for the status of mental states (especially propositional attitudes). These issues should be of considerable interest to social scientists whose ontology and explanatory framework is based on the existence and causal efficacy of propositional attitudes. Having argued (in chapters 2 and 3) that phenomenal experiences (henceforth PEs) are an epiphenomenal consequence of events in the brain, the debate surrounding connectionism and eliminative materialism provides a forum within which we can assess the causal efficacy of the content of mental states. I will conclude part I by fleshing out the version of epiphenomenalism that is developed in the first four chapters and anticipate and reply to some potential objections.

It has been claimed that if epiphenomenalism turns out to be true that it will be the most important revolution in our world-view to date. Mankind may no longer see itself as being God's creation or as sitting in the centre of the universe, but we certainly think of ourselves as different in kind from any other physical system. We are conscious and we believe that our consciousness allows us to escape the causal chain that determines the behaviour of all other physical systems (except, perhaps, at the quantum level). This assumption is the bedrock of all that is distinctively human, if we choose to abandon it then we have to re-examine everything that philosophy and the social sciences have to teach us. Moral responsibility is the first and most obvious casualty, but our sense of personal identity is also brought into question along with our assumptions regarding how we should go about explaining people's behaviour. Emotions such as love, honour and duty are robbed of their virtue. Creativity and original thought can no longer be viewed as an achievement of their author (defined as a conscious subject), since they become, like all other behaviour, merely the result of blind physical forces. Also called into question is the justification for our holding certain rational arguments to be true. If a theory or conjecture is merely the result of neuronal processes, and there is no rational subject to evaluate their worth, what justification do we have for judging them to be true or false?

Examining the full consequences of epiphenomenalism would take several lifetimes, let alone being within the scope of a single thesis. Nevertheless, certain issues emerge quite naturally out of the discussion in part I and these will be explored in part II. I attempt, in part I, not only to argue in favour of epiphenomenalism, but try to get clear the relationship between conscious events and the physical events that cause them. Doing so then allows us to question such things as the relationship between the self and the brain, how conscious rational thought is related to non-conscious physical events and how explanations of behaviour that are based on the assumption of mental causation (folk psychological explanations) should be treated. Although the discussion of such subjects will necessarily be rather sketchy they do, I hope, provide at least the basis for more research and some suggestions as to how the social sciences (and the rest of humanity) might come to terms with epiphenomenalism.

Before diving in at the deep end of the debate on mental causation I want to take this opportunity to introduce some concepts and clear up some of their ambiguities. The following section is, I'm afraid, a little tedious, but so much hangs on precision and conceptual clarity in this debate that a bit of tedium is unavoidable.

Supervenience and the metaphysics of causation

Despite first receiving an enthusiastic reception in the philosophy of mind it now looks like the tide is turning against supervenience. Those theorists who once thought that supervenience might provide a useful tool in the search for a theory of mental causation now worry that supervenience may render all macroproperties causally impotent. Before we can begin to consider mental causation, therefore, we had better ensure that we are working with a metaphysics of causation that allows for the causal efficacy of macroproperties. By macroproperties I have in mind all those properties that are cited by the special sciences. To put it another way, all physical properties that are *not* at the level of the most basic particles and their properties. What we are after is a theory of causation that does not entail that properties such as being a spark plug, being fragile, or being a calculator are epiphenomenal. This section is designed primarily as an introduction to the issues we will face when confronting supervenient mental causation. I will, however, also be discussing issues relating to causation in general that will be of relevance throughout the thesis. I will begin by providing a brief introduction to the concept of supervenience before outlining Kim's theory of epiphenomenal and supervenient causation. I will then take a brief look at the role causal laws play in explanations before concluding this section by considering the curiously titled concept of quausation.

If the mental supervenes on the physical then a molecule for molecule identical copy of me would also be psychologically identical. If aesthetic properties supervene on physical properties then there could not be two works of art identical in all physical respects but differing in some aesthetic characteristic. Similarly, if moral characteristics supervene on one's actions then there could not be two men who face identical circumstances, and behave in an identical manner, but who differ in their moral attributes.[7] In its simplest form supervenience states nothing other than a pattern of property covariance between a supervenient property or properties (such as mental states, aesthetic properties, or moral virtues) and a subvenient base (such as a person's neurophysiology, the brushstrokes of a painting or the actions of a man). Notice that in highlighting this pattern of property covariance supervenience tells us nothing about what causes the instantiation of a given higher order property and nor does it make that supervenient property reducible to or identical with its subvenient base. What the concept does do, however, is to ground certain higher order properties firmly in the physical world. I should stress at this point that the philosophical usage of the term supervenience differs from its meaning in everyday English. In everyday

[7] The latter use of supervenience was developed by R. M. Hare who introduced the concept to denote the relationship between goodness and the set of properties or qualities that constitute goodness. According to Hare, if we take St Francis to be a good man, it is logically impossible that a man in the same circumstances as St Francis, who behaved in exactly the same way, is not also a good man (Hare 1952: 145).

parlance to supervene is to occur subsequently, or as a change, interruption, or addition. The philosophical usage of the term implies no time lag or alteration and is far from synonymous with addition. As will become clear, supervenience is a relation between two synchronous properties rather than two discrete events or states.

There are now several formulations of the supervenience thesis in the offing – strong versus weak and regional versus global being the main parameters. Which version of the thesis one chooses depends largely on the properties in which one is interested. Moral or aesthetic properties are likely to require a wider subvenient base than causally potent properties and mental properties are likely to have a still narrower subvenient base (though this is debatable). We will touch on some of these issues later, particularly with reference to the causal role of content bearing states. As far as possible though, I will try to avoid getting bogged down with the purely logical/conceptual debates and focus on supervenience as applied to causation and mental causation. Throughout this thesis, unless stated otherwise, when I use the term supervenience I will be referring to strong supervenience (defined as follows):

> A *strongly supervenes* on B just in case, necessarily, for each x and each property F in A, if x has F, then there is a property G in B such that x has G, and *necessarily* if any y has G, it has F. (Kim 1984a: 65)

I will use Kim's example of 'being a good man' to illustrate this relation. For our purposes it would be better if we could illustrate this example using a physical property, the subvenient bases for physical properties are, however, considerably more difficult to pin down. Properties such as being a spark plug, being fragile, or being a calculator, can not only be realised by a vastly larger number of subvenient bases, but since they are defined in part by the application to which they are put the subvenient base would have to include relational properties.

Moreover, if we were to illustrate strong supervenience with a physical property we would have to reckon with the mereological problem of identifying its subvenient base. If we went down to the microphysical level, and there are good reasons (to be discussed later) for supposing that we may have to, it would become impossible to fit a description of the subvenient base, if such a description were available, into a single volume. Kim's example then will have to suffice.

Suppose that the property of being a good man supervenes on the character traits of honesty, courage and benevolence. If St Francis is a good man then there is some combination of these traits (say honesty and benevolence) that St Francis has, and necessarily anyone who has these traits is also a good man. The combination of traits that St Francis has is obviously not the only combination of traits that together warrant the description 'being a good man'. If St Francis had lacked honesty but been courageous, for example, he might still have been a good man. One important point that is not illustrated by Kim's example is that supervenient properties have taxonomic priority. Supervenience does not allow us to first identify a subvenient base and then infer the existence of supervenient properties. We can only make such an inference once we already know that a given higher-order property *A* is realised by a physical base *B* and we have encountered a physical system *S* that is micro-indiscernible with respect to *B*. Only then can we use our knowledge of the microphysical properties of *S* to infer the presence of *A*. Any difference in the microproperties of *B* and *S*, no matter how seemingly trivial, precludes our making this inference. Kim's illustration of being a good man is slightly misleading in this respect since it appears to suggest that if a person *x* has the properties of honesty and benevolence that *x* too would necessarily be a good man. This is not the case though since if *x* were misogynistic and cruel to animals as well as being honest and benevolent he might not qualify for the description 'being a good man'.

To summarise supervenience:

(i.) states a pattern of property covariance,

(ii.) implies an asymmetric dependency relation (in that it does not make sense to say that the character traits of honesty and benevolence supervene on being a good man),

(iii.) is consistent with the multiple realisation of supervenient properties by a range of physical bases,

(iv.) gives taxonomic priority to the supervenient property,

(v.) requires indiscernibility of subvenient bases to guarantee both instantiate the supervenient property.

With the concept of supervenience now in place we can begin to approach the issue of causation where the causal properties are macroproperties. Here we will look at Kim's theory of epiphenomenal and supervenient causation. One of Kim's most instructive examples used to highlight what he means by epiphenomenal causation concerns the relationship between a disease and its symptoms. About this Kim says, 'the symptoms are not mutually related in the cause-effect relationship, although to the medically naïve they may appear to be so related. The appearance of a causal connection merely points to the real causal process underlying the symptoms' (Kim 1984b: 94). Epiphenomenal causation then, refers to cases where there is a real cause-effect relationship occurring at one level and an apparent casual relationship occurring as a consequence at some higher level. In this thesis I argue that PE is analogous to the symptoms of a disease and that mental causation (where mental is read as referring to PE) is a species of epiphenomenal causation. That is to say, the PEs of, for example, a thirst followed by a desire for water, are related in a real cause-effect relationship, but the relationship is at the neurological level and the experiences themselves are, like the symptoms of a disease, not mutually related in a cause-effect relationship.

Kim later, and rather confusingly, extends the use of the term epiphenomenal causation to cover cases of supervenient causation where there is a real causal relationship between macroproperties. Supervenient causation is thus a subclass of epiphenomenal causation. In what follows I will retain the use of the term epiphenomenal causation to refer exclusively to those illusory relations exemplified by the relationship between a disease and its symptoms. In the next chapter we will then examine whether the physicalists' application of supervenience succeeds in demonstrating the existence of genuine supervenient causation or just epiphenomenal causation. Unlike epiphenomenal causation, supervenient causation expresses a real causal relationship between two macroproperties or events in the following way: 'x's having F supervenes on x's having m(F), y's having G supervenes on y's having m(G), where m(F) and m(G) are microproperties relative to F and G, and there is an appropriate causal connection between x's having m(F) and y's having m(G)' (ibid. 99). Thus, if the supervenient cause of desiring a drink of water is the experience of thirst, the relation is as follows: Kim's having the PE of thirst supervenes on m(T), his having a desire for water supervenes on m(D), where m(T) and m(D) are the microphysiological events that realise the PE of thirst and the desire for water respectively, and there is an appropriate casual connection between Kim's having m(T) and m(D). Without some account of the relationship between m(F) and F, and m(G) and G, of course, there is nothing to differentiate supervenient from epiphenomenal causation. For the purposes of this chapter we will set aside the job of accounting for the relationship when the macroproperties are mental (as in the thirst-desire example), and focus on cases where the macroproperties are physical. What differentiates the two theses is the causal role played by the microphysical subvenient base of the macroproperties involved. In a case of epiphenomenal causation, such as the progression of a disease and its symptoms, the microphysical event that is the subvenient base(s) for the symptoms is not causally efficacious in the progression of the disease. In a case of supervenient causation, such as a fatal blow to the head, the microphysical event(s) that is the subvenient base for the fatal blow is causally efficacious in causing death.

For Kim supervenient causation (of macroproperties) is ubiquitous: 'All observable phenomena are macrophenomena in relation to the familiar theoretical objects of physics; hence... all causal relations involving observable phenomena – all causal relations familiar from daily experience – are cases of epiphenomenal causation' (Kim 1984b: 95).[8] The ubiquity of supervenient causation means that unless the causal relations between macroproperties are real (that is to say they are related as cases of supervenient causation and not just epiphenomenal causation) no causal explanations would be possible. If the causal relations aren't real then every scientific explanation, every counterfactual, and every scientific law, expresses at most an observed regularity or constant conjunction as Hume would say. Since one could not possibly hope to give a

[8] Remember that for Kim supervenient causation is a species of epiphenomenal causation. It is supervenient causation that Kim is referring to here.

microphysical explanation of macro-events, one would have to give up all hopes of causal explanation.

Lynne Rudder Baker argues that strong supervenience (henceforth SS) and the principle of the causal closure of the physical (henceforth CCP) entails just such a scenario. In response to Kim's theory of supervenient causation Baker argues that SS plus CCP means that the following relevance condition should be true:

> If an instantiation of a microphysical property m(F) is a complete cause of an instantiation of micro-physical property m(G), and necessarily, for every instantiation of m(F), there is an instantiation of F, and necessarily, for every instantiation of m(G), there is an instantiation of G, then the instantiation of F is causally relevant to the instantiation of G. (Baker 1987: 88)

According to Baker the trouble with Kim's account, or indeed any account that accepts CCP and SS, is that causally related microproperties stand in the same relation to non-causally related macroproperties as they do to causally related macroproperties. Baker hopes that the following illustration will encourage her readers to join her in concluding that SS plus CCP leads to a *reductio ad absurdum*:

> suppose that a person exercising in front of a mirror jumps up and comes back down. Let m(F) be the micro-physical properties of the space-time region that includes the mirror and the exerciser as she goes up, m(G) be micro-physical properties of the space-time region that includes the mirror and the exerciser as she comes down; let F be some macro-properties of the reflection as the exerciser goes up, and G be some macro-properties of the reflection as she comes back down. Necessarily anything that has m(F) has F, and necessarily, anything that has m(G) has G. Therefore, by [the above relevance condition], the instantiation of the properties of the reflection in the mirror as the exerciser went up caused the instantiation of the reflection's properties as she came down. (ibid. 89)

As we shall see this sceptical conclusion is unwarranted. The relevance condition has F supervening on *m*(F) and G supervening on *m*(G). It does not, however, begin from the premise that F causes G. Rather, it states explicitly that *m*(F) causes *m*(G). Most cases of causation involve both epiphenomenal and supervenient causation so it is not in the least surprising that *one* of the properties which supervenes on *m*(F) is related as an instance of epiphenomenal causation to *one* of the properties which supervenes on *m*(G). Such cases of epiphenomenal causation are well known to science and indeed are utilised on a daily basis by people such as general practitioners when they enquire as to their patients' symptoms. The relevance condition fails to state that F and G are the *only* properties instantiated by *m*(F) and *m*(G). It merely states that the instantiations of F and G are causally relevant – which of course they are. An observer that has a view of the exerciser's reflection, but not of the exerciser, could use their observations of the exerciser's reflection going up to predict that the exerciser would soon come back down. Notice though that in the illustration employed by Baker her language as shifted from 'causally relevant' to 'caused'. Such a shift in emphasis is entirely unwarranted since it would only be valid (by SS, CCP, and the relevance condition) if the instantiation of the reflections in the mirror of the exerciser going up and down were the *only* properties instantiated by *m*(F) and *m*(G). Since *m*(F) and *m*(G) include all the microphysical properties belonging to the time-space region that contain both the exerciser *and* the mirror, this is patently not the case.

Baker does note that one way to avoid her sceptical conclusion is to claim that Kim's account was not intended to allow any inferences about macrocausation from microcausation. Instead Kim's account might be intended as a means of allowing us to give an account of microcausation for a given case of macrocausation. In response to this potential objection Baker appeals to the exclusion argument claiming that 'since the problem was that CCP and SS seemed to leave no room for macro-causation, one needs a reason to believe that macro-causation exists' (ibid. 89). I want to argue that Baker has failed to show that macroproperty epiphenomenalism is entailed by SS plus CCP. The possibility of being able to provide a complete causal explanation at the microphysical level, plus the assumption that macroproperties supervene on microproperties, has absolutely no implications for the causal powers of macroproperties. To get macroproperty epiphenomenalism from SS plus CCP one needs to add some further premises regarding the relationship between macroproperties and their subvenient bases. If one is happy to accept, as I am, that there is no real ontological depth to the reductionists' layered world view – that causal powers belong to events and that macro and microphysical descriptions are descriptions of the same events at different levels of analysis – then there are no implications for macrocausation to be derived from SS plus CCP. What one needs to get macroproperty epiphenomenalism from SS plus CCP is some premise to the effect that macroproperties are *caused*[9] by microproperties. Materialism neither advocates nor entails such a view.

Nevertheless the exclusion argument does become important for the causal powers of mental properties. Here the ontological relationship between mental events and their subvenient base is uncertain so the explanatory exclusion of the mental is a real possibility. Kim formulates the exclusion problem as follows: '*Given that every physical event that has a cause has a physical cause, how is a mental cause also possible?*'(Kim 1998: 38) In other words, if it turns out that mental states supervene on neural states, and if a neural state *N* at *t* contained sufficient conditions for a later event *E*, it would be a contradiction to say that a mental state *M* that supervened on *N* at *t* contained necessary

[9] By caused I have in mind the common sense understanding of a cause being distinct from its effect.

conditions for *E*. Neither, of course, could *M* contain sufficient conditions for *E*, since this would result in overdetermination. Here again, however, the formulation of the exclusion argument requires the addition of further premises. If it turns out that token-token physicalism is true (that every token mental event is identical to a token physical event) then it makes no sense to ask 'how is a mental cause also possible?' It makes no sense, of course, because the mental cause would be the same event as the physical cause. Similarly, there would be no contradiction in saying that *N* contained sufficient conditions for *E*, and *M* contained necessary conditions for *E*, if *N* and *M* are the same event under two different descriptions (where *N* refers to all the properties and powers belonging to an event and *M* is a more restrictive description of the same event). The explanatory exclusion of the mental only becomes a possibility if one adds a premise to the effect that *N* and *M* are two different events (or so I shall argue). Much of the next chapter will be concerned with establishing whether P1 (the irreducibility of mental states) is a strong enough ontological claim to entail the explanatory exclusion of the mental given SS and CCP. This is an issue that, with the notable exception of Kim (see Kim 1993d), is rarely discussed in the contemporary literature on mental causation. As I hope to show, however, it is still the most important problem facing contemporary theorists of mental causation and it is a problem which has been obscured rather than eliminated by talk of supervenience.

Though explanatory exclusion may be a problem for theorists of mental causation there are no good grounds for believing that it extends to macrocausation generally. Fodor diagnoses all those who have succumbed to this philosophical worry with a case of epiphobia. Epiphobia, according to Fodor, is a neurotic worry that stems form two philosophical mistakes '...(a) a wrong idea about what it is for a property to be causally responsible, and (b) a complex of wrong ideas about the relation between special-science laws and the events that they subsume' (Fodor 1990). Fodor's account is concerned with the role of properties in causal laws and their relation to the events that instantiate them. In what follows I will present a simplified model of Fodor's argument and argue that, although it provides causal relevance for some special-science properties, it hangs on a theory of reference that fails to secure causal efficacy for mental properties (especially PE).

Fodor's argument is based on the following assumptions (all of which are relatively unproblematic):

1. *Covering principle*: If an event *e*1 causes an event *e*2, then there are properties *F*, *G* such that:
   1.1. *e*1 instantiates *F*
   1.2. *e*2 instantiates *G*

and

   1.3. "*F* instantiations are sufficient for *G* instantiations" is a causal law.
   2. P is a causally responsible property if it's a property in virtue of the instantiation of which the occurrence of one event is nomologically sufficient for the occurrence of another. (ibid. 143)

What Fodor's account hangs on is the relationship between events and the laws that *cover* or *subsume* them (Fodor's phraseology) and how the law '...*projects* the properties in virtue of which the individuals [events] are subsumed by it' (ibid. 143). Fodor claims that on his view whether a property is causally responsible *reduces* to the question of whether there are causal laws about that property. Although this is trivially true, in the sense that if we accept that the physical world is both casually closed and law governed it must be the case that causally efficacious properties are subsumed by causal laws, as a criterion for identifying causally efficacious properties it fails. In order to demonstrate this it is necessary to examine what Fodor means by basic and non-basic causal laws.

On Fodor's account, a law is basic if the properties projected by that law could not be reduced. That is to say, that there are no further-to-be-explained facts about the properties. A law is non-basic if there is a story to tell about *how* the properties projected by a law are causally efficacious. To take Fodor's example 'meandering rivers erode their outside banks' is a non-basic law since there is a microstructural story to tell about the effects of particles suspended in the water. The law relating to these microproperties would be a basic law.

Fodor is right that the question of whether a property is causally responsible reduces to the question of whether there are causal laws about that property. However, a strong case can be made that this is only the case when the laws are basic. The trouble is that it is possible for epiphenomenal properties to be related in a lawlike way without being causally responsible. Suppose, for example, that there is a law relating to hay fever sufferers that states that itchy eyes cause the eyes to water. Suppose also, for the sake of the argument, that this law is exceptionless. Thus, by Fodor's argument, itchy eyes are a causally responsible property in virtue of the fact that the occurrence of an event which instantiates itchy eyes is nomologically sufficient for an event that instantiates watery eyes. It still remains a possibility, however, that itchy eyes are related to watery eyes as an example of epiphenomenal causation. The only way that Fodor's account can be modified to prevent epiphenomenal properties slipping through is to reformulate 1.3 of the covering principle such that it refers to microphysical laws. In the case of properties invoked by a non-basic law, we are only justified in calling those properties causally responsible if they *refer* to micro-structural properties that are *projected* by a basic law. To explicate, we are justified in calling 'being a recessive trait' (one of Fodor's examples) a causally responsible property because it is a property projected by a non-basic law *and* because it accurately refers to its implementing mechanism (DNA). A problem occurs if the property projected by a non-basic law has a multiply realisable microstructure (subvenient base) with disjunctive causal powers. For example, it is *the property of being a mountain*, according to Fodor, that causes Mt. Everest to have snow on top. Not so, the property of being a mountain has a multiply realisable microstructure with disjunctive causal powers (geologists refer to mountains and valleys on the seabed which clearly do not share the same causal powers as Mt. Everest).

Notice that this appeal to causal laws as a means of identifying causally potent macroproperties does *not* state that macroproperties are causally efficacious *because* they figure in casual laws. Davidson's truism that causation is a relation between concrete events no matter how they are described is all too easily forgotten in these days of epiphobia (see the latter half of Davidson 1980). As Davidson rightly quips: 'Naming the American invasion of Panama "Operation Just Cause" does not alter the consequences of the event' (Davidson 1993: 8). Figuring in causal laws may help us to identify causally efficacious properties and justify our attributing causal powers to them, but we do not need to know the relevant law to identify the cause of an event.

Quausation

Although Fodor's account, if modified along the lines suggested above, does guarantee the causal powers of *F* it does not entail the more specific claim that *F* quaused *G*. The concept of quausation (i.e. c *qua F* causes e *qua G*) was developed by Terence Horgan (whose account of mental quausation I will discuss in the next chapter). Here though I will illustrate the concept of quausation with LePore and Loewer's (1989) account:

> <*c,F*> is quausally related to <*e,G*> iff[10] *c* and *e* occur and are respectively *F* and *G* and there is some time before the occurrence of *c* at which these two conditionals obtain: (1) if *c* were to occur and be *F* then that would cause an event *e* to be *G*; (2) if *c* were to occur and not be an *F* then it would not cause an event which is *G*. (LePore 1989: 189)

Suppose, to take a simple example, I drop a stone into a still pool thereby creating ripples. The ripples are caused by the stone displacing a certain volume of water and we can say with absolute confidence that dropping the stone caused the ripples. Quausation, however, is a more specific concept than causation and while one can say with absolute confidence that the ripples were caused by dropping the stone we can say with equal confidence that the ripples were not quaused by dropping the stone. It was not the stone *qua* stone that caused the rippling but the stone *qua* object with a specific mass, volume, shape, hardness, etc. That is to say, if we had dropped a relevantly similar object (i.e. one with the same mass, volume, shape, etc.) in place of our original stone the causal relations at issue would be unaffected.

Now let's consider a case where a property is quausally efficacious. Suppose Whistler's 'Old Battersea Bridge' were to come up for sale at auction. Let *c* be the event of Old Battersea Bridge coming up for sale and *e* be its sale, let *F* be the property of being the original painting and *G* be the event of its being sold for *x* amount. Quausation allows us to single out the property of 'being the original' as the quausally relevant property relative to *G* as follows: the property of being the original painting is quausally related to its eventual sale price iff *c* and *e* occur and are respectively *F* and *G* and there is some time before the

[10] Iff is the abbreviation for 'if and only if.'

occurrence of *c* at which these two conditionals obtain: (1) if the original were to come up for sale it would cause the sale price to be *x*; (2) if *c* were to occur and not be the original (say, for example, the painting that comes up for sale is a known fake indistinguishable from the original) then it would not cause the sale price to be *x*.

Quausation then is really just a sophisticated counterfactual that helps to clarify the relationship between a property in a cause-event and a property in an effect-event. The trouble with quausation is that it cannot cope with cases of epiphenomenal causation where the epiphenomenal properties are necessarily instantiated by the cause-event and the effect-event. To explicate, suppose that a disease necessarily instantiates its symptoms and the symptoms strongly supervene on the microphysical properties of the disease. In such cases, if we attempted to apply the counterfactual expressed by quausation it would show that the symptoms supervening on the cause-event (say the characteristic ache that accompanies the onset of flu) are quausally related to the symptoms supervening on the effect-event (headache, sore throat, etc.). It may be possible to rectify this weakness by clarifying the relationship between *c* and *F*, and *e* and *G*. I will say no more about this here but we will return to this issue in the following chapter when I discuss Horgan's account of mental quausation.

To summarise, we have identified three different types of causation: epiphenomenal causation, supervenient causation and quausation. These different types, it should be noted, are not mutually exclusive. If a property is implicated in a given causal transaction: as epiphenomenal causation it may not also (by my restrictive definition) be related as an instance of supervenient causation or quausation; as supervenient causation it may also be, but is not necessarily, a case of quausation but cannot also be a case of epiphenomenal causation; and as quausation it is necessarily related as a case of supervenient causation but cannot also be a case of epiphenomenal causation. Finally it should be noted that all three forms of causation confer causal relevance on the property in question but only quausation and supervenient causation provide scope for causal efficacy. I should point out here that my use of the term causal relevance is not synonymous with casual efficacy. As I use the term a property is causally efficacious if it is literally the cause of, or a partial cause of, a later event. A property is causally relevant if its instantiation is useful for predictive purposes. In the next chapter we will consider three approaches to mental causation that make use of supervenience. In each case we will first attempt to establish which type of causation the supervenience theorist has in mind and whether they are successful. I hope to show that, given P1-3 (premises which are shared by most of the authors to be considered) the PE of mental events may only be related as cases of epiphenomenal causation. The content of mental states, in contrast, may be related as cases of supervenient causation and may also be quausally efficacious. Having shown that the causal efficacy of the content of mental states is consistent with our three premises, however, is a long way from demonstrating empirically that they are causally efficacious. This latter question will be considered in chapter 4 where I will discuss connectionism and eliminative materialism.

The most important point to be carried over into the next chapter is that neither counterfactuals nor causal laws are able to demonstrate the causal efficacy of macroproperties unless the relationship between macroproperties and the microproperties upon which they supervene (or from which they emerge) has been clarified. This, as we shall see, is even more vital when the macroproperties involved are mental.

Chapter 2

Mental Causation: Supervenient Causation or Epiphenomenal Causation

By now we have a fairly clear idea about the various ways that macroproperties enter into causal relations. We have yet to consider, however, what is meant by *the mental* in mental causation. At the outset I stressed the importance of PE for the orthodox mental realists' concept of mental causation. It seems to me that any theory that denies that my itching *literally* caused my scratching or that my wanting a glass of wine *literally* caused me to open a bottle (or at least asserts that these aspects were necessary elements in a total cause) fails to qualify as a theory of mental causation. To be clear on this point, we would not expect any theory (materialist or dualist) to assert that PE is a sufficient cause of action.[1] Rather, we would expect PE to be relevantly engaged in a causal chain that involves such things as nerve impulses, muscle contractions, neurophysiological events and the like. Moreover, if mental events strongly supervene on neurophysiological events, we would not expect the neurophysiological base of a given mental event to be a sufficient cause of action. Instead what we would expect is for the neurophysiological base to be a necessary part of a total neurophysiological event that is itself a sufficient cause for action (*ceteris paribus*, of course).

With the above in mind we can agree with Davidson's (1980b: 221) definition that an event is mental if and only if it has a mental description (i.e. contains at least one mental verb). This definition makes intentionality the mark of the mental. Typically intentionality is associated with a PE – there is something it is like to believe, admire, think, wish etc. Though many people would argue that unconscious mental states display intentionality, these are surely parasitic on conscious mental states.[2] A theory of mental causation then would be one that explains how one event with a mental description causes either a second event, which again has a mental description, or an action. Moreover, the theory must explain how the mental aspect of an event is a necessary condition for producing its effects. As we shall see, Davidson's own theory of anomalous monism arguably fails to fulfil the latter criterion.

Anyone who is tempted to apply the supervenience thesis to mental causation has to perform a rather tricky balancing act with epiphenomenalism on one end and dualism on the other. The difficulty arises because physicalism requires that there be a tight connection between mental and physical properties. The tighter the connection the more kosher the theory is from a physicalist perspective. Ideally, the physicalist wants a theory that ensures a strict dependency relation between mental and physical properties. The trouble is that a strict dependency relation entails that physical properties alone become sufficient for causal explanation. There is no work left for mental properties and we are left with either a case of overdetermination (a violation of P3) or epiphenomenalism and the explanatory exclusion of the mental. On the other hand if the dependency relation is relaxed, thereby making room for mental causation, we risk violating the principle of the causal closure of the physical and end up with dualism – Descartes' revenge as Kim (1998: 38-47) terms it.

[1] With the possible exception of actions some dualists might count as wholly mental, a decision, mental calculation, etc. I will not complicate matters here by considering such actions or indeed whether such events should count as actions at all.

Part of what makes getting to grips with mental causation so difficult is that the term mental is used in an infuriatingly ambiguous manner. Although I think that most people would agree that a theory of mental causation should provide an ineliminable role for PE, this is a tall order to live up to. There is, therefore, a strong and understandable temptation to be a little vague when outlining a theory's explanandum. Consequently one often finds theorists hinting that their theory will secure a causal role for PE only to discover that the causal role of content, the function of mental states or the subvenient base of mental states is their real (and often unstated) target. I would not wish to diminish the importance of these latter projects but it is important to be clear about the explananda of the theories we are about to consider. After pinning down the explanandum we will have to question the type of causation that the theory secures (epiphenomenal causation, supervenient causation, quausation, or causal relevance) and assess whether any of our three premises have been violated in the process.

In what follows I will consider three approaches to mental causation that arguably have different explananda. (i.) That the mental is casually efficacious in virtue of its phenomenal properties. (ii.) That the mental is causally efficacious in virtue of its physical properties or subvenient base. The debate surrounding this version centres on whether this is enough to secure causal efficacy for mental states or just causal relevance. (iii.) Finally, there is the (broadly) functionalist strategy of claiming that the mental is causally efficacious in virtue of its content. Typically this version does not go into detail about how content is instantiated. For each version of mental causation we will examine a paradigmatic account:

[2] See Searle, J. (1992) ch. 7

for (i.) we will look at Horgan's account of quausation; for (ii.) Davidson's anomalous monism and Kim's multiple-type physicalism; for (iii.) Dretske's dual-explanandum strategy. I should note that each account is considered heuristically for the light it can shed on theories of its type rather than for its own merits.

Mental Quausation

I have already outlined the concept of quausation in relation to macrophysical properties. There I argued that quausation may prove to be a useful tool in identifying the causally efficacious properties from the substantial set belonging to a given event. However, when those properties are macroproperties I argued that one needs to be clear about the relationship between the macroproperty in question and its subvenient base. Without getting this relationship clear there is a danger that epiphenomenal properties relative to the instance of causation in question (such as the symptoms of a disease) may slip through and count as quausally relevant. Now Horgan subscribes to a form of the token-identity thesis where every token mental state is identical to a token physical state that guarantees the causal efficacy of mental states. Thus, if I were to drop a stone on my foot and subsequently shriek with pain, we can say with absolute confidence (given token physicalism) that the pain caused my pain behaviour. However, just as in the case of macrocausation discussed in the introduction, even assuming token physicalism is true, quausal epiphenomenalism remains a possibility. That is to say, the event/state that is picked out by the description 'the pain that resulted from dropping the stone' caused my behaviour, but it remains possible that it was not the pain qua pain (i.e. the PE of pain) but the pain qua electro-chemical event/state that caused my behaviour. If it was the electro-chemical event/state that caused my pain behaviour then token physicalism notwithstanding the PE of pain is quausally epiphenomenal. This anyway is how I interpret the problem Horgan seeks to tackle.

Although quausal epiphenomenalism (of the PE of mental states) is a real possibility, indeed I am arguing in favour of what might be termed a version of quausal epiphenomenalism in this thesis, it is not a real possibility within the framework of token physicalism. The reason is that token physicalism denies that events, such as those that occurred in my head following my dropping the stone on my foot, are composed of two distinct sets of properties, one mental and the other physical. For token physicalism pain is not composed of the PE of pain (phenomenal pain for short) plus the physical instantiation of that pain (or physical pain). Rather, pain is a single event/state that can be described using two different vocabularies. As Davidson has persuasively argued, causation is a relation between concrete events no matter how they are described, thus when Horgan claims that quausal epiphenomenalism is a problem even for token physicalism he is confusing the ontological question of what pain is with the epistemological issue of how that pain is described. Nevertheless, Horgan's account is still worthy of consideration within the context of nonreductive physicalism which does admit the existence of two distinct sets of properties.

Finally, before delving into Horgan's account, we still have the job of identifying which aspect of mentality his account is designed to secure. As I have previously argued, the most natural interpretation of the mental in mental causation is PE. PE is, after all, arguably the only aspect of mentality that is necessarily mental (the content of mental states, though dependent on consciousness, may be unconscious). Thus we could interpret Horgan's project as an attempt to save the PE of mental states from being reduced to the status of epiphenomena. At times though it looks like Horgan is going for the more modest task of securing a causal role for the content of mental states. We will consider, in what follows, quausation applied to both these projects. I will argue that the former task cannot be achieved without ditching one or more of our three principles (something that neither Horgan nor myself would be prepared to do). If his task is the more modest one of securing a role for the content of mental states then Horgan fares better. A victory in this latter task, though of philosophical importance, is of little consolation for those that view epiphenomenalism as an affront to human dignity. The causal efficacy of content, though necessary for the causal efficacy of PE, is not sufficient.

Horgan tells us that quausation is designed to ward of the possibility of quausal epiphenomenalism. Traditional epiphenomenalism, the view that mental states have no effects at all, is summarily dismissed on the basis that token mental states are identical to token physical states. Quausal epiphenomenalism is defined as a position that denies that '... mental events and states *are* causes ... and that they have the effects they do *because* they instantiate the specific mental properties they do' (Horgan 1989: 51). Thus, according to Horgan, unless it can be shown that the mental *qua* mental is efficacious, quausal epiphenomenalism remains a possibility. Indeed, we are told that this is the most pressing problem in contemporary philosophy of mind. Horgan's account runs as follows:

c *qua* F causes e *qua* G if:

(i.) event c causes event e,
(ii.) c and e respectively instantiate properties F and G,
(iii.) F and G are logically and metaphysically independent, and
(iv.) the transaction between c and e does not involve pre-emption, overdetermination, or the like,

then the fact that c and e instantiate F and G respectively, is explanatorily relevant to the fact that c causes e iff the following *Relevance Condition* is satisfied:

(R) For any world w in P[c,e],[3] if c* is the event in w that is pertinently similar to c of W[4], then

(i.) if c* instantiates F in w, then c* causes (in w) an event e* which both instantiates G (in w) and is pertinently similar to the W-event e; and

(ii.) if c* does not instantiate F in w, then event c* does not cause (in w) an event which is pertinently similar to the W-event e. (Horgan 1989: 58/9)

The first four conditions are consistent with properties F and G being epiphenomenal; the real work of rooting out causally efficacious properties is done by the relevance condition. Though the relevance condition is designed to show how properties F and G can be quausally efficacious, it merely formulates (albeit in a philosophically rigorous way) a pattern of counterfactual dependency. In other words, it tells us that if the event were to occur again and the causally efficacious property were not present, then the effect would be different. So, for example, in the much discussed case of a man killed by a loud gunshot, if the bullet were replaced by a blank the relevance condition would show that sound was not the efficacious property: if c* does not instantiate F (the property of being a real bullet) in w, then c* does not cause (in w) an event which is pertinently similar to the W-event (there is a loud bang but it does not kill the man). Thus, the relevance condition purports to show that it is the property of being a projectile, not a loud bang, which kills the man. Horgan notes (ibid. 59) that this pattern of counterfactual dependency is often sought empirically. Quite how it could be used to demonstrate the causal efficacy of mental states remains a mystery – how does one go about removing mentality to check whether it is causally efficacious?

[3] P[c,e] refers to pertinently similar worlds (PSW) where the correlates c* and e*, as well as the background conditions w, resemble in pertinently similar intrinsic respects c and e, and the background conditions in W.

[4] W is the actual world in which event c causes event e.

The real worry, however, is whether quausation is compatible with both supervenience (either strong supervenience or Horgan's own theory of regional supervenience outlined below) and the token-identity thesis to which Horgan adheres. Horgan accepts that all our actions are explainable, in principle, with reference to the laws of physics as applied to our microphysical parts (see Horgan 1989: 64). We are also told, however, that whether or not event c instantiates property F is causally relevant to its effect. More precisely, whether or not event e instantiates property G is dependent on whether or not event c instantiates property F. If we interpret the phrase 'instantiates' as referring to supervenience, then F supervenes on c and G supervenes on e. The trouble is that supervenience asserts a strict dependency relation between a supervenient property and its subvenient base such that necessarily anything that has the subvenient base also has the supervenient higher order property. This strict dependency relation poses something of a problem for the relevance condition, since it entails that necessarily if c instantiates F in W, c* must also instantiate F in w.

This means that, for the mental to be causally efficacious qua mental, a token mental state F cannot be identical to a token physical state c. If F were identical to c then F's causal powers would be identical to those of c and, therefore, it would make little sense to say that F is quausally efficacious (since this would be equivalent to saying that c is quausally efficacious). Moreover, if F really were quausally efficacious (because it possessed some causal powers over and above those of c) then not only must the assumption of token-identity be mistaken but the principle of the causal closure of the physical would be violated– a position which, as we have already seen, Horgan accepts. On the other hand, if it turns out that the mental is not causally efficacious qua mental and the token-identity thesis is true then the causal powers of F would be identical to those of c and we would have a violation of P1 (the irreducibility of mental states).

Horgan's concept of regional supervenience does provide one way out of this impasse. As we shall see, however, it does so only for relational properties and so cannot be used to argue in favour of the causal efficacy of mental states when the mental is read as referring to PE. Horgan's account of regional supervenience runs as follows:

> There are no two P-regions that are exactly alike in all qualitative intrinsic physical features but different in some other qualitative intrinsic feature. (Horgan 1993: 571)

Horgan's point is that the subvenient base of a higher order property *F* need not be intrinsic to the individual that instantiates the higher order property – it need not supervene on 'what's in the head'. Indeed Horgan makes a persuasive case for widening the subvenient base of certain higher order properties beyond synchronous, internal events and states – the wide-content mental property *wanting a drink of water* is a case in point (Horgan 1993: 571). Suppose, for example, to borrow one of Horgan's examples, that Oscar has a sudden desire for a glass of water. According to Horgan, Oscar's having this property is not 'an intrinsic feature of the spatio-temporal region directly occupied by Oscar's body' during the period in which Oscar desires water. In simple terms Horgan is claiming that Oscar's desire for water (a wide-content property) does not supervene on what's in Oscar's head. Nevertheless Horgan thinks that this wide-content property is casually implicated in Oscar's behaviour. The reason for this is that the property's instantiation is dependent upon Oscar's having acquired the term 'water' in a world where water is H_20 rather than XYZ. For Horgan then, strong supervenience is not strong enough to guarantee the causal-explanatory relevance of wide-content. Perhaps the reason that Horgan feels that he needs to appeal to regional supervenience is because he has run together two different explananda. Oscar's desire for water is constituted, in part, by a PE of thirst. This PE, assuming PE supervenes on synchronous internal physical states, is an intrinsic feature of the spatio-temporal region occupied by Oscar's body. If the PE of thirst is a sufficient condition (*ceteris paribus*) for Oscar to take a drink then all the causal-explanatory work is done by phenomenal states plus their subvenient bases that are intrinsic to Oscar. Whether Oscar's desire is for H_20 or XYZ is causally irrelevant in relation to his subsequent behaviour.

To recap, the thrust of this section has been concerned with demonstrating that quausation is incompatible with supervenience. The trouble is that no matter how one interprets the *qua* relation as soon as one starts attributing causal powers to the quausal property (the F in c *qua* F) the supervenience relationship between c and F no longer holds. For F, where F is some PE, to have any quausal powers it must be the case that F is a new property added to c or a property generated by c but distinct from it. Either way it looks like quausation is straying into the territory of emergentism.

Mental causation as physical causation

While many would wish for a theory of mental causation that has the mental qua mental doing the causal work, for authors like Davidson (arguably) and Kim it is the mental as physical that is causally responsible. Although, as it is sometimes put, causal efficacy is 'transmitted' from the physical to the mental in virtue of the supervenience relation, this approach is open to the charge of epiphenomenalism on the grounds that mental properties are related as epiphenomenal causation rather than supervenient causation. We will begin with Davidson's Anomalous Monism (henceforth AM). As the instigator of the debate on mental causation in its current form, the literature on AM is voluminous. In this brief critique we can barely scratch the surface of the complex issues that surround AM. Nevertheless, as I noted in the introduction, when it comes to what I consider to be the failure of theories of mental causation the devil has not been in the detail. As such we can afford to skate over some of the trickier issues and focus on the crucial question of the relationship between mental and physical properties. Let's begin by outlining Davidson's approach.

AM states that all events are physical, though some events are also mental, and denies that mental phenomena can be given purely physical explanations. AM was designed to reconcile an apparent contradiction between the following three principles. That mental events interact causally with physical events. That events related as cause and effect fall under strict deterministic laws (the Principle of the Nomological Nature of Causality). And that there are no strict laws on the basis of which mental events can be predicted and explained (Davidson 1980b: 208). We will start by exploring these three principles.

The first principle states the familiar assumption that mental events interact causally with the body and, via the body, with the external world. Our first worry concerns how this interaction is to be conceived. Davidson holds a version of the identity thesis where token mental events are identical to token physical events. Moreover, Davidson conceives of causation as a relation between concrete events no matter how they are described. It is difficult to imagine, therefore, what is being claimed by the premise that mental events interact causally with physical events that could not be expressed by 'physical events interact causally with physical events'. Indeed, as Kim points out, this premise does not adequately capture our common sense understanding of what it means for a mental event to be causally efficacious. For, as Kim says, in the context of AM 'the claim that "Mental events cause physical events" only comes to the assertion… that events with some mental property or other are causes of events with some physical property or other' (Kim 1993a: 20). It is consistent with (though not implied by) AM, therefore, that events with epiphenomenal mental properties cause events with physical properties. This is clearly not what Davidson has in mind. What we need is some account of how the mental properties of a cause-event are causally efficacious in producing the effect-event. This, however, does not seem to be available within a Davidsonian ontology of event causation. We will return to this issue after outlining the second premise in more detail.

The second premise is the lynchpin of AM. The Nomological Character of Causality states that where two events are related as cause and effect there must be a strict law covering the case. A strict law in this case refers to the type of laws that one would expect to find in a finished physics. Causal laws, for Davidson, relate to (and due to the nature of causal laws, must relate to) events as described and not individual dated events. Davidson suggests, for example, that '… it would make no sense to speak of an individual event being "invariably accompanied" by another' (Davidson 1980b: 212). As we shall see, it is Davidson's belief that mental kinds cannot be correlated with physical kinds that supports his theory of the anomalous nature of the mental. Having noted that the Nomological Character of Causality does not refer to determinism, this should not be taken to mean that Davidson is an indeterminist. Indeed Davidson does appear to endorse determinism. In *The Material Mind*, for example, Davidson describes a *l'homme machine* (called Art) that has been built in man's image according to a finished physics, chemistry, neurobiology, etc. All of Art's physical behaviour, including his speech acts and intentional behaviour (physically described), would, according to Davidson, be entirely predictable. Despite this Davidson maintains none of Art's mental events (so defined) could be predicted or explained (neither, I suspect, would Davidson claim they could be retrodicted). The reason for this, and the reason why Davidson views the mental as anomalous, is that mental and physical schemes are incommensurable:

> It is a feature of the mental that the attribution of mental phenomena must be responsible to the background of reasons, beliefs, and intentions of the individual... we must stand prepared, as the evidence accumulates, to adjust our theory in the light of considerations of overall cogency: the constitutive idea of rationality partly controls each phase in the evolution of what must be an evolving theory. (Davidson 1980b: 222-3)

Davidson concludes that if we conceive of man as a rational animal then we must accept the nomological slack between the mental and the physical (this is, of course, the third principle).

Subsequent discussion has focussed on whether the above three premises are consistent with AM and the concept of supervenience 'taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect, without altering in some physical respect' (Davidson 1980b: 214).

It is the role played by psychophysical laws that has been most widely debated. Many commentators have claimed that psychophysical laws are required to *fix* the mental to the physical. Kim states the problem as follows: '...*the very same network of causal relations would obtain in Davidson's world if you were to redistribute mental properties over its events any way you like; you* *would not disturb a single causal relation if you randomly and arbitrarily reassigned mental properties to events, or even removed mentality entirely from the world.* The fact is that under Davidson's anomalous monism, mentality does no causal work' (Kim 1993c: 269).

Let me digress briefly, I have previously endorsed Davidson's extensionalist view of causal relations as relations between concrete particulars no matter how they are described. I agree with Davidson, therefore, that 'it makes no literal sense... to speak of an event causing something as mental, or by virtue of its mental properties, or as described in one way or another' (Davidson 1993: 13). Nor, of course, would it make literal sense to speak of an event causing something as physical, or in virtue of its physical properties. Nevertheless, it seems intuitively obvious that some of the properties belonging to an event are causally efficacious while others are epiphenomenal (relative to a given causal transaction). Consider, for example, Sosa's (1984) illustration of the loudness of a shot (epiphenomenal relative to causing death) and the property of being a projectile (causally efficacious in causing death). The trouble is that if we agree that it would make no sense to speak of an event causing something as mental, or as a projectile, can we then consistently maintain that PE is epiphenomenal? In other words, if it makes no literal sense to speak of an event causing something as PE or in virtue of its phenomenal properties described in one way or another then it makes no literal sense to claim that the opposite is true – that an event did not cause something in virtue of its phenomenal properties, or as described in some way or another. We will return to this issue later.

Kim's argument is that on AM mental events are causally efficacious in virtue of their falling under physical types and that a mental type tokened by a given mental event is epiphenomenal. That is not to say that Kim is contradicting the premise that mental events cause physical events (no-one, myself included, has charged Davidson with this type of epiphenomenalism). Rather, it is Kim's assertion that under AM events are causes only *as* they instantiate causal laws. This, Kim maintains, means that mental properties are epiphenomenal because 'to suppose that altering an event's mental properties would also alter its physical properties and thereby effect its causal relations is to suppose that Psychophysical Anomalism, a cardinal tenet of anomalous monism, is false' (Kim 1993d: 270). Davidson has objected to this line of criticism (Davidson 1993: 6) on the grounds that given his extensional views on causation, as I quoted above, it makes no literal sense to speak of an event causing something as mental or by virtue of its mental properties. In reply to Davidson, Kim notes (Kim 1993a: 20) that his claim was that mental *properties*, not mental *events*, make no causal difference under AM. Terminological quibbles aside, however, the problem still remains that AM seems to leave no causal role for mental properties. Here things start to get a little technical but it is worth spending some time ironing out these problems since, as I noted above, if an extensional concept of causation precludes talking about the causal powers of properties (mental or physical) then the epiphenomenalism I advocate would seem to make no literal sense either.

Brian McLaughlin has made a number of cogent points on this issue. In particular he provides an account of why Davidson thinks that his extensionalist view of causation (C1) is incompatible with C2:

> C2 If event *c* caused event *e*, then *c* caused *e* in virtue of certain of *c*'s properties. (McLaughlin 1993: 31)

According to McLaughlin, Davidson 'thinks that the claim that event *c* causes event *e* in virtue of *c*'s having *F* implies that *c*'s having *F* causes *e* (or that *c* causes *e* under the description 'the *F*'). And since C1 implies that causes are events, C1 and C2 are, he thinks, incompatible' (McLaughlin 1993: 33). Far from being incompatible, McLaughlin argues that C2 actually implies C1. To explain why, McLaughlin turns to the 'weighs less than' relation discussed by Davidson (1993: 6):

> That *a* weighs less than *b* in virtue of weighing 10 pounds, does *not* imply that *a*'s weighing less than 10 pounds weighs less than *b*. A's weighing less than 10 pounds is a state of affairs, and states of affairs have no weight. Thus, to be sure, *a*'s weighing 10 pounds does not weigh less than b; for that is nonsense. But if *b* weighs (say) 11 pounds and *a* weighs 10 pounds, then *a* weighs less than *b* in virtue of weighing 10 pounds. And this implies that *a* itself weighs less than *b*. Moreover, if *a* weighs less than *b*, then *a* does so however *a* and *b* are typed (or described). (McLaughlin 1993: 33-4)

The same goes, McLaughlin argues, for an event's having a certain property F. Thus the claim that event *c* causes event *e* in virtue of *c*'s having *F* does not imply that the state of affairs of *c*'s having *F* causes *e*. This is good news for our purposes since, it seems, we can consistently maintain an extensionalist view on causation while at the same time maintaining that PE is causally inefficacious. It does make literal sense, in other words, to claim that event *c* causes event *e*, event *c* instantiates property *F* (where property *F* is some PE), but that event *c* does not cause event *e* in virtue of *c*'s having (instantiating) property *F*.

The identity of mental events with physical events, therefore, is insufficient to guarantee the causal efficacy of mental properties. The supervenience relation in this case does guarantee that mental properties are causally relevant, and that they are related to physical properties (and other mental properties) by a form of epiphenomenal causation, but it fails to show that mental properties participate in causal relations as supervenient causation. Ultimately AM faces the same problems as Horgan's quausation considered earlier. It seems inevitable that any version of nonreductive physicalism that subscribes to the view that mental properties supervene on physical properties will face the same fate. One simply cannot consistently maintain that PE is irreducible, supervenient on physical states, and causally efficacious, without violating either P2 or P3. The identity of mental and physical events is not enough, what materialist theorists of mental causation require is the identity of mental properties with physical properties. Any such identity would, of course, be a violation of our first principle.

Multiple type-physicalism

Unlike many other authors in the field, Kim is sensitive to the distinction between the causal efficacy of mental states defined in terms of their PE, and the causal efficacy of their subvenient base. This is a rather difficult distinction to get one's head around, and matters are not made any easier by Kim's conspicuous lack of examples. One cannot blame Kim for this since the possibility of real life examples is precluded by the supervenience relation. To explicate, remember that SS states that a given subvenient base will invariably be accompanied by its supervenient property (or properties). Thus, if PE strongly supervenes on the physical, there are no circumstances where one could have (for example) the neural correlate of pain without the experience of pain. It therefore makes no sense to ask how a person with a severe headache would behave if they lacked the neural correlates of pain. And neither would it make any sense to ask how a person would behave if they had the neural correlates but experienced no headache. This is essentially the problem that quausation tried, and I argued failed, to solve.

Like most authors (the only notable exceptions being the eliminativists), Kim is keen to secure a place for mental states in causal explanations. Unlike most authors, however, Kim seems (rather reluctantly) willing to give up the efficacy of the PE of mental states in favour of the content of mental states as realised by physical states. One can discern in Kim's work a growing acknowledgement that one cannot save the causal efficacy of PE without falling foul of the exclusion problem and thereby making events overdetermined (violating P3), or violating the physicalist principle of the causal closure of the physical (P2).

Multiple-type physicalism states that mental properties supervene on disjunctive physical bases. To explicate, that the same mental property, say pain, may be realised in different systems, by a variety of physical bases – P_h in humans, P_m in molluscs, and P_M in Martians. Further, that P_h may be realised by different physical bases in different individuals, in different circumstances, or even in the same individual at different times in their life. 'In this sense, we may say that mental kind M is *disjunctively identified* with physical kinds P_1, P_2,... Note that M is not identified with the *disjunction* of P_1, P_2...; nor is an M-instance identified with an instance of the disjunctive property $P_1 \vee P_2 \vee$... We may call this proposal "multiple-type physicalism"' (Kim 1993e: 364). Thus, for an event or object to have a mental property M, is for M to be physically realised by one of its physical bases. Moreover, M's causal powers are identical to those of its physical base. This makes '...mental properties causally inhomogeneous in the sense that two instances of the same mental property may have quite diverse causal powers, and that the more diversely a mental property is realized, the greater its causal inhomogeneity' (ibid. 362). One of my aims in this thesis is to demonstrate that not all instances of a given mental state *M* are equally efficacious and to develop some sort of taxonomy or framework for identifying the relative causal efficacy of the subvenient base of *M* vis-à-vis other synchronous neural states. As such Kim's realisation model provides a sound philosophical basis for such an endeavour.

One of the strengths of Kim's thesis is that, not only does it escape many of the problems related to mental causation that we considered above (causal overdeterminism and the explanatory exclusion of mental states), but that it does not try make philosophy of mind the final arbiter of the causal efficacy of mental states. In short, it leaves science (cognitive, computational and neurological) with a real job to do. It does not attempt to settle the question of what, or how wide, the physical base of mental states are, and consequently leaves open the question of their causal efficacy.

In order to demonstrate this lets take an action (A), 'opening a bottle of wine', motivated by a token conscious reason (M), 'I fancy giving that new Pinot Noir a try'. On Kim's view, M's causal powers must be identical with a token of a physical kind, P_1, P_2... (whichever realised M). The crucial point is that on Kim's model how wide P_1, P_2... turns out to be is left open. Thus if it turns out (it doesn't) that M is realised by P_1, and P_1's causal powers are identical with the whole brain and nervous system of the organism in which it is instantiated, then we can say that M caused A. Alternatively it might turn out that P_2 realised M and that P_2 is identical with the firing of a few dozen neurons somewhere in the brain that are neither necessary or sufficient for A. In which case M does not cause A.

Multiple-type physicalism attempts to solve the problem of mental causation by functionalising mental states, or, more specifically, functionalising the content of mental states. This requires that we adopt an all-encompassing reductionism about the mental and entails that mental states have no new causal powers over and above those of their physical realisers. Although functionalising of the content of mental states does appear to be the only means of solving the problem of mental causation that is compatible with physicalism and supervenience, as Kim is acutely aware, it does not work for the PE of mental states. Functional properties are causally efficacious in virtue of the role they play within a system. The phenomenal qualities that a property possesses are consequently irrelevant to its being able to perform its function. This is the antithesis of the orthodox interpretation of the causal efficacy of mental states where the PE is causally efficacious precisely because it has specific phenomenal qualities. If qualia are causally efficacious in virtue of their function then there seems no good reason why, for example, inverted qualia[5] should not be possible. Because the PE of mental states resists functionalisation, Kim is led to the following dilemma:

> Ultimately we are likely to face the following choice: either embrace the realization view and save mental causation, or insist on the unique and distinctive status of mental properties, especially the qualia, but be prepared to give them up as causal powers. The paradoxical thing about this is that the choice offered may only be the illusion of a choice, for the two options may in the end collapse into one. If you choose the former, you may loose what makes the mental distinctively mental; and what good is it, one might ask, if you save mental causation but end up losing mentality in the process? ... If you choose the latter, again you may lose the mental, for what good is something that is causally impotent? ... Being real and having causal powers go hand in hand. (Kim 1993e: 366/7)

The loss of mentality resulting from the former choice is the belief that we, as conscious subjects and because we are conscious subjects, have control over our

[5] Examples of inverted qualia might be hot and cold, pleasure and pain, red and green etc.

own behaviour. The realisation view strips us of this power by reassigning the causal power of mental states to the brute and unthinking forces of the physical domain. We lose mentality twice over on this view. On the first scenario, by functionalising mental states and hence saving the causal efficacy of their content, the distinctive PE of mental states is robbed of its causal role. That is to say, from a functionalist perspective what matters is the function that the subvenient base plays in a given causal chain, not the characteristic PE of the supervenient state. On the second scenario, where PE can't be functionalised, we are left with only two choices, we must either adopt PE irrealism (which is not an option) or claim that they are irreducible hence epiphenomenal. Kim seems to think that these alternatives amount to the same thing since he views having causal powers as an essential criterion for being real. From a methodological perspective this is quite true, since by definition something entirely devoid of causal powers could never be discovered (being perceivable, by for example showing up on a scientific instrument, constitutes a causal power). Thus from a pragmatic perspective Kim is quite correct that having causal powers is an essential criterion for being real. PE, it seems plausible to suggest, escapes this criterion because it is necessarily experienced. Where causation is a relation between two or more events, a given instance of PE does not need to cause anything in order to make its presence literally felt. Moreover, the reason that we have concluded that the PE of mental states is epiphenomenal is precisely because they are *sui generis* real and irreducible to physical states. The reality of phenomenal states is indubitable; it is only their relationship to physical states that is in question.

The realisation view is also likely to face difficulties because of its close association with functionalism, which many philosophers find rather unpalatable. Although there are certain similarities, what differentiates Kim's model from classic functionalism is that on a functionalist account causal efficacy derives solely from the function that a given property plays within the system in which it is instantiated. Functionalists are typically not interested in how the property is realised and will allow a property to be realised by a potentially infinite set of systems. Thus, functionalists will frequently allow a property *M* to be instantiated by, and have the same causal powers in, systems as diverse as computers, biological systems or the population of China.[6] On Kim's model, on the other hand, because *M* strongly supervenes on a subvenient base *P*, we can set limits on the physical systems that can instantiate *P*. Which physical systems are capable of instantiating *M*, therefore, becomes an empirical rather than logical or conceptual matter.

Of all the supervenience theories on offer multiple-type physicalism is the only one to secure a form of mental causation while respecting the irreducibility of PE and without violating either P2 or P3. By adopting a broadly functionalist approach Kim succeeds in securing the causal efficacy of the content of mental states. We may lose an aspect of mentality (PE) from mental causation on this account but it is the argument of this thesis that this is an inevitable consequence of the three principles outlined in the introduction. For this reason multiple-type physicalism, or something like it, must be the correct way to characterise the relationship between mental and physical properties. In all likelihood Kim's theory will be conceptually refined and empirical

[6] See (e.g.) Searle (1992) and Block (1978) for a critique of functionalism.

observations will certainly tell us about the conditions necessary for a system to realise a mental state. Nevertheless, if the argument presented in this thesis is correct then it would take a paradigm shift in the philosophy of mind to overturn Kim's basic argument – namely only the successful reduction of PE or the repudiation of the causal closure principle will overturn his theory.

Dretske's dual-explanandum strategy

Thus far I have argued (following Kim) that real causation only occurs at the most basic level – whatever that might turn out to be – and that macrocausation has to be considered as supervenient or epiphenomenal causation. Moreover, that macroproperties can only be considered causally potent if they are identical to microproperties (though we need not know what these microproperties are). Though providing an ontological account of causation is an essential foundation for a sound methodology, it is of little or no practical use in itself. What we require is an account of causation that is sensitive to these ontological considerations, but that can form part of a methodology capable of prediction and explanation. That is, one that allows us to individuate the salient properties involved in causation without having to give a microstructural account of an event and all its antecedent conditions.

Fred Dretske provides just such an account in his *Reasons in a World of Causes* (1988). His dual-explanandum strategy provides a useful account of how to individuate the salient causal properties from amongst an event's antecedent conditions. In doing so, Dretske persuasively argues that reasons, and their cognitive and conative components, rationalise (or causally explain) behaviour. Though I will argue that Dretske does not solve the problem, his approach is worth considering because, unlike many theorists, he is acutely aware of the problem of causal explanatory exclusion. That is to say Dretske, as a physicalist, is willing to admit that a physical explanation, citing only the causal connections between the most basic particles, combined with the most fundamental physical laws and concepts, is sufficient to explain and predict all our movements without invoking intentional states. Yet, according to Dretske, reasons, beliefs and desires do causally explain our behaviour. His project is to reconcile these seemingly incompatible principles.

I will begin by outlining Dretske's distinctions between behaviour and movement, and triggering and structuring causes, before moving on to consider his naturalised account of meaning and his suggestion as to how meaning may figure in causal explanations of behaviour. Here I will draw on Kim's account which clears up some of the confusion relating to the causal role of token and type indicating states and makes explicit the supervenience relation which is implicit in Dretske's account.[7] As ever we will consider the role-played, on Dretske's account, by the PE of mental states and the type of causation that his account proposes.

Dretske begins by distinguishing between behaviour and movement. Behaviour is to be construed as endogenously produced movement; it is something the organism does rather than something that is done to the organism (Dretske 1988: 2). Dretske is emphatic that we must not conflate the motor output *M* of an organism *S*, which has an internal cause *C*, with the cause of an organism's behaviour. When we enquire as to the cause of an organism's motor output, we are asking the how question: *how* does *C* cause *M*? When we ask,

[7] Dretske (1991: 216) endorses Kim's formulation of his account and acknowledges that it elucidates the type-token distinction.

'what caused *S*'s behaviour?' we are asking a *why* question: why did *C* cause *M*. Put simply, physiology explains how *C* caused *M*, reasons explain why *C* caused *M*.

Dretske's dual-explanandum strategy thus avoids the problem of overdetermination by claiming that physiology and psychology do not share the same explanandum. On Dretske's model the why question can be further subdivided:

> In looking for the cause of a process we are sometimes looking for the triggering event: what caused the *C which* caused the *M*. At other times we are looking for the event or events that *shaped* or *structured* the process: what caused *C to* cause *M* rather than something else. The first type of cause, the triggering cause, causes the process to occur *now*. The second type of cause, the structuring cause, is responsible for its being *this process*, one having *M* as its product, that occurs now. (Dretske 1988: 42)

Reasons causally explain behaviour, according to Dretske, by explaining the structuring cause (the cause of *C*'s causing *M*). In the case of a classically trained dog, for example (see ibid. 43-4), the ringing of a bell produces an auditory experience (C) in the dog, which subsequently causes the dog to salivate (*M*). We may then ask what caused the dog to salivate and can answer by providing either a triggering cause, the sound of the bell ringing, or a structuring cause, that the dog has learnt to associate the ringing of a bell with food.

To give a full account of the structuring cause in this case we would have to explain how the auditory experience of hearing a bell causes the dog to believe that food is on its way. Beliefs, on Dretske's account, are internal representations or *maps* that help us steer through the world and that figure in the cause of movement. In addition though, in order to count as a belief:

> *The fact that it is a map*, the fact that it *says* something about external conditions. must be relevantly engaged in the way it steers us through these conditions. What is required... [see fig 2.1 below] is that the structure's indicator properties figure in the explanation of its causal properties, that what it *says* (about external affairs) helps to explain what it *does* (in the production of output). (Dretske 1988: 94)

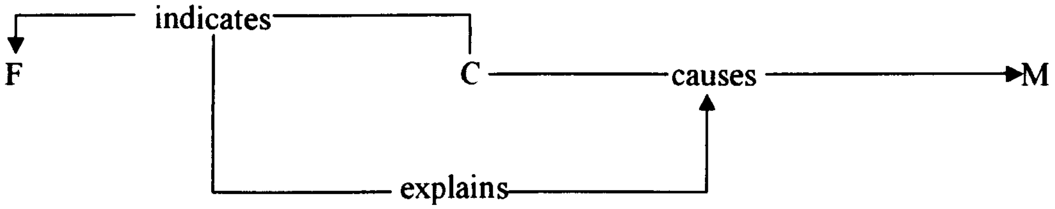


Fig. 2.1 The role of indicator properties in Dretske's dual-explanandum strategy (taken from Dretske 1988: 84)

Organisms acquire structures with indicating properties, on Dretske's model, through the familiar process of associative learning. Thus, if we return to the example of the classically trained dog, the bell ringing causes *C* which (due to associative learning) has the property of indicating *F*. *C*'s indicating *F* (the dog's belief that food is on its way) causally explains why *C* causes *M*.

Kim explains how reasons (R), on Dretske's account, explain behaviour in the following way:

> *Stage* 1: Why does C's having R at t cause M?
> Because C's having R at t supervenes on C's having N [nonrelational, intrinsic neural property] at t, and there exists in S an N → M causal structure.
> *Stage* 2: But why is the N → M structure present in S? How did content property R (i.e., the property of representing F) come to supervene on N in S? And how did S acquire the capacity to represent F?
> Because, in S, N is the internal indicator property of F, and through a process of conditioning and learning the N → M causal structure was established in S; this means also that in S, R came to supervene on N. In the process, N has acquired the function of indicating F, and this is how S came to have the capacity to represent F. (Kim 1991: 68)

One of the strengths of Dretske's account is that it provides a causal role for wide content while respecting the assumption that the triggering cause of behaviour, to use Dretske's phrase, must be internal and virtually synchronous with the movement it is held to explain. That wide content does not supervene on 'what's in the head' is now almost received wisdom in philosophy.[8] If wide content is to figure in causal explanations of behaviour, therefore, it must do so

[8] See (e.g.) Putnam (1975)

by causally influencing the triggering cause. Dretske develops an information-based account of meaning that allows meaning to do just that. Very crudely the meaning of an internal state *C* derives from its function of indicating *F* – it carries information about (e.g.) the presence of *F*. This makes meaning dependent on extrinsic environmental conditions: *C* represents *F* because there is a reliable *correlation* between the occurrence of *F*, and *C* representing *F*. Moreover, *C*'s having meaning *now* is dependent on extrinsic and historical environmental factors. It is because *C* *was* reliably correlated with *F* during the learning process that results in its meaning (indicating/referring to) *F now*.

It is not clear exactly what Dretske means by meaning. His approach is designed to give the semantic character of brain structures (or meaning) a real job to do. He does not want to make them epiphenomenal consequences of syntactic structures. He is quite emphatic about this and gives us the example of a soprano who, when singing a passage with a specific meaning, shatters glass due to the pitch of her voice. The meaning of the passage is clearly epiphenomenal; though it may be causally relevant if it turns out that only those phrases with a particular meaning reach the causally potent pitch. Dretske notes that: 'If having a mind is having *this* kind of meaning in the head, one may as well not have a mind' (ibid. 80). However, he later goes on to say that: 'Whatever else a meaning might be, it certainly is not, like an event, a spatio-temporal particular that could cause something to happen. No, in exploring the possibility of a causal role for meaning one is exploring the possibility, not of meaning itself being a cause, but of a *thing's having meaning* being a cause or of the *fact that something has meaning* being a causally relevant fact about a thing' (ibid. 80). In the end though Dretske's account of the causal efficacy of a thing's having meaning boils down to the function that internal indicating mechanisms play. The only thing that differentiates a semantic structure from syntactic structure is the causal history of the indicating structure in question. I have to admit that I am at a loss to understand why Dretske places such importance on this distinction.

The trouble with this account of meaning, and this is probably true of most attempts to naturalise meaning, is that it completely sunders meaning from the PE of the meaningful. We, as systems with intrinsic (or original) intentionality, *feel meaning*; systems possessing derived intentionality (even computational systems capable of 'learning') have no such feeling. Though we may be said to believe many things that we are never conscious of, when we are conscious of our beliefs there is a PE of believing, and this PE seems to be missed out of naturalised accounts of meaning.

The point is nicely illustrated by Horgan who asks us to imagine:

> ... a Frankenstein creature, created by Martians whose mastery of robotology and human neurophysiology is vastly superior to ours. The Martians deliberately design the robot with neural circuitry very much like the circuitry that might have been instantiated by a sophisticated and well informed philosopher, in 1990 America, who has an enormous amount of knowledge about the world but has total amnesia about his own past.... Imagine that the creature is activated in Dretske's presence, and immediately begins arguing vociferously... that it already has full-fledged propositional attitudes... (Horgan 1991: 87)

Horgan argues that Dretske is committed to denying, because this robot has not yet learned anything such that his beliefs could be endowed with content, that it does not act on its beliefs or that it even has beliefs. This example illustrates what has been my primary concern in this chapter, the role of the PE of mental states. Suppose that the Martians had constructed the robot with identical (molecule for molecule) neural circuits to an actual 1990s American philosopher, since PE is supervenient on the synchronous internal physical states of the system in which it is instantiated, it must be the case that the robot has exactly the same experiences as the original philosopher. (At least at the moment of activation, thereafter unless the philosopher and his copy were in identical environments their experiences would differ.) There is a real tension between defending the claim that the robot has no beliefs and the physicalist's requirement that the robot experiences having beliefs. The only way out of this problem is to relax the requirement that content bear the right causal history to what it denotes.

One way to do this is to draw a distinction between reference and meaning. The content of the robot's thoughts do not refer to the objects/events they 'seem' to remember (since they have no causal connection to those objects/events). However, whatever internal (innate) structures allow the robot to use its native language are not robbed of their meaning because they fail to refer to the objects/events which they denote. Both innate and learned indicating structures have potentially the same causal powers. They are both capable of misrepresentation (something upon which Dretske places great emphasis), and both could potentially provide the subvenient base for PEs. Compare, for example, two different species which each have a fear of humans, but in the first case the fear is innate and in the second, learned. In both species the neural mechanisms that subserve this fear have the same casual powers (they cause the animal to flee at the sight of a human, for example), are capable of misrepresentation (they may flee at the sight of a scarecrow) and both are capable of generating the PE of fear with the appropriate stimuli.

Dennett also seems puzzled about Dretske's denial that innate mechanisms can figure in *why* explanations. Dennett suspects that Dretske has been distracted by the illusion that in the case of learning, but not of natural selection, 'the organism itself (or even: its mind or soul) does the understanding

– responds directly to the meaning' (Dennett 1991b: 123). In natural selection the 'understanding' is done by the brute forces of selection, there is no understanding left over for the organism to do. If this is the reason why Dretske rejects natural selection then, as Dennett notes, 'of course, if "mere conditioning" is responsible for the redesign of the individual organism's brain, this too looks like taking responsibility away from the inner understander that Dretske is so loath to lose' (ibid. 124).

To return for a moment to the distinction Dretske draws between structuring and triggering causes, Dretske has made a persuasive case that reasons causally explain behaviour when they are cited as the structuring cause. However, when reasons are cited in the context of a structuring cause they do not relate to the PE of mental states (indeed PE need not be a cited at all when reasons relate to a structuring cause). They are, in effect, third person descriptive accounts of why the causal *mechanisms* became hooked up in the way that they did. This focus on mechanisms, and the functions they perform, has the effect of making PE irrelevant for the purposes of explanation. Though mental states are of more relevance to the triggering cause, again it is the function that a given mental state plays that is of relevance not its PE. Thus Dretske succeeds in demonstrating that reasons can causally explain behaviour, but fails to show that I (as conscious subject rather than object of study) perform action A because I have a conscious reason R. We will encounter this distinction again in the next chapter when we consider Searle's theory of mental causation.

Summary and conclusion

Supervenience, it will be remembered, originated as a method of characterising moral and aesthetic properties so as to show that they are not something extra, something in addition to the actions of a man or the brush strokes of a painting. This characterisation of property covariance works well for the relationship between a mental property and its subvenient base. It should be obvious from the foregoing discussion that, as applied to causation, supervenience is being used to assert something more than property covariance. In arguing in favour of mental *qua* mental causation one is surely claiming that mental properties are something in addition to physical properties. The burden of causation is more than the concept of supervenience can handle. Inevitably, it seems, whenever mental properties are characterised as both supervenient properties and causally efficacious properties we run into problems such as overdetermination or the violation of the causal closure principle.

The supervenience thesis also encounters serious difficulties when it is removed from the relative safety of pure philosophy and applied to empirical observations. There is now a considerable research interest in exploring the role consciousness plays in cognition, perception, memory storage and retrieval, information processing and voluntary acts. Thus far, though the evidence is by no means conclusive, there appear to be at least some instances where conscious awareness results from these processes but does not directly enter into them. Although we are far from the stage when the neurosciences can pronounce on the issue of mental causation versus epiphenomenalism, we now have good reason to believe that at least some of the processes to which we had attributed consciousness a central role are performed unconsciously (see Velmans 1991). Conscious awareness, in these processes, appears to be a consequence of causally efficacious neural states rather than their cause. Even if it turns out that these processes are the exception rather than the rule, philosophical accounts of supervenience must be able to account for them. Supervenience, and supervenient causation, are presented as general theories of the relationship between mental and physical events, if there are mental processes where consciousness can be shown to be an epiphenomenal consequence of physical events, supervenience theorists can not just dismiss them as anomalies. In the remainder of this chapter I will consider one of the putative cases of epiphenomenalism that has emerged from the neurosciences and explore whether or not the concept of supervenience can account for it. I should make it absolutely clear at this point that I am not attempting to use empirical material to argue in favour of epiphenomenalism. Indeed, for the purposes of the following discussion, it does not matter much whether the following putative case of epiphenomenalism turns out to be true. The point of the following discussion is merely to highlight the difficulties that arise when the 'supervenient cause' (PE of deciding to initiate an action) is not temporally coextensive with the 'subvenient cause' (neurophysiological events).

One of the best documented, and most discussed, cases of epiphenomenalism concerns the feeling of 'voluntary' control over the flexion of one's finger. In an intriguing set of experiments Benjamin Libet et al set out to establish the position of conscious awareness and decision making in the causal chain that results in the flexion of one's finger. To do so they recorded cerebral activity, the time of a subject's decision to act and the time movement was first initiated. By using surface electrodes to measure the onset of a readiness potential (henceforth RP), and directly comparing RP with the time of the subjects conscious 'wanting' or intending to act, Libet et al were able to demonstrate that RP preceded movement by an average of about 800 ms for the onset of the main negative component, or by 500 ms taking the onset of RP at 90 per cent of surface area (Libet et al. 1983: 630). The conscious awareness of the decision to act occurred around 150-200 ms before the onset of movement. Libet et al conclude that cerebral initiation of a freely voluntary act can begin unconsciously and several hundred milliseconds before the subjective awareness of a decision.

Libet et al do not go as far as claiming that consciousness is epiphenomenal because, they claim, it is still possible that consciousness has the power of veto over unconscious 'decisions'. This veto option seems unlikely though since, as Spence (1996) and Velmans (Velmans 1991) have noted, neuronal states must persist for at least 400-500 ms before neuronal adequacy is reached.[9] Thus, any conscious veto would also result from neuronal states – in effect, a conscious veto would be the epiphenomenal consequence of an unconscious veto.

This putative case of epiphenomenalism point to some difficult issues for supervenience, supervenient causation and token physicalism. Typically, concepts of supervenience make higher order mental properties supervene on *synchronous* internal physical states. Further, they assert that token higher order properties, and their causal powers, are identical with their corresponding token physical state. Upon what physical state is the conscious decision to act supposed to supervene? The problem is that the neural event to which we attribute causal efficacy (the RP) and the conscious 'decision' differ in duration. The neural event precedes the conscious event by some 300-600 ms (depending upon

[9] Neuronal adequacy refers to the temporal duration and intensity neuronal states must reach before conscious awareness is generated (see Libet 1993)

whether one takes the main negative component or RP at 90 per cent of surface area as the event's cause) which means that the PE of the conscious decision lasts only 25%-50% as long as the RP. In order to be consistent with any of the theories of supervenient causation considered above, we would have to identify the PE of the conscious decision to act with only those neural events with which it is temporally coextensive. The problem here is that the neural events with which it is temporally coextensive can only be of a 150-200 ms duration (which is the time between the first occurrence of the conscious decision and the initiation of movement) as we have already noted though, neuronal adequacy requires 450-500 ms. Thus, in order to make supervenient *causation* compatible with this empirical material, one would have to identify a higher order mental property with a *preceding* neural state. Identity of this kind would be incompatible with the temporal order of causation as it is currently understood. In effect it would endorse backwards causation. It seems that there is going to be no neat fit between token mental states (supervenient properties) and token neural states (their subvenient base).

Chapter 3

Emergence

Life emerged from a rich and varied molecular soup. The complex behaviour of an ant colony emerges from a simple set of genetically prescribed rules that govern the behaviour of each ant. Consciousness emerges from a few pounds of neurons, glial cells and the rest. I emerged from my tent to watch a glorious sunrise. The economy is an emergent property of individual agents but is irreducible to the behaviour of individual agents. Other than indicating some basic change of state, these various usages of the term emergence have very little in common. Emergence is such an evocative term that its wide spread use is to be expected, however its application in contexts as varied as sociology, biology, physics, philosophy and the popular lexicon has created a quagmire of conceptual confusion and ambiguity. Moreover, with one or two notable exceptions, most authors are inexcusably lax in their characterisation of the term (thereby fuelling further cross-disciplinary confusion). Before we can get to the interesting bit, therefore, we have a bit of tedious (though ultimately fruitful) definitional and classificatory work to do. The pay off for this work is that it will greatly simplify our critique of the various theories of emergence. As I see it there are four distinct senses of the term emergence: emergent properties, emergent patterns, emergent complexity, and emergent powers. It is the claim that consciousness exhibits emergent powers that is of most relevance to our present inquiry since it poses a threat to epiphenomenalism. Though most of our attention will be focussed on emergent powers I will also spend some time exploring the potential application of the other versions.

Before discussing these four types of emergence I want to briefly consider and set aside the issue of predictability which is frequently invoked as a criterion for emergence. It is often said that a property or power is emergent if its occurrence could not have been predicted.[1] Predictability is an ontological red herring because whether or not the occurrence of a higher order property is predictable depends on (1) whether or not we can identify its constituent parts. And (2) whether there are general laws and theories available regarding the behaviour of its parts in isolation and combination. In the present context (whether or not one can predict the occurrence of conscious mental events based on knowledge of lower level physical processes), the second condition is only approximately satisfied. We suspect that the laws of classical physics and possibly those of quantum mechanics, govern the behaviour of brain. The first condition is no where near being satisfied. In the case of consciousness it is not clear what parts we should be investigating: plausible candidates include neurons and sub-neuronal structures, neuronal assemblies, neurotransmitters, brain regions, brain regions with particular frequencies, the list could go on. Moreover, although we have now amassed a considerable body of knowledge concerning the operation of individual neurons and small 'assemblies' of neurons, we have virtually no idea how the brain performs even simple cognitive tasks. That predictability if not a genuine ontological trait was recognised by many of the early emergentists, so why contemporary advocates cling to non-predictability as the criterion for emergence is something of a mystery. Hempel summarises the point as follows:

> ...emergence of a characteristic is not an ontological trait inherent in some phenomenon; rather it is indicative of the scope of our knowledge at a given time;

[1] See, for example, Popper and Eccles (1977).

> thus it has no absolute, but rather a relative character; and what is emergent with respect to the theories available today may lose its emergent status tomorrow. (Hempel 1965: 263)

In the context of the present debate, prediction is a weapon that belongs in the hands of the reductionist camp. For if it can be shown that the effects of a putative emergent property are predictable based solely on the 'bottom up' causation of its constituent parts (more on this later), then it will have been shown that the emergent property does not possess emergent powers. However, absence of such evidence should not be used, especially at this early stage in research, as evidence of emergent powers (explained below).

Emergent properties

There are two senses of the term emergent properties, one trivial and one with real ontological depth. The trivial sense refers to a property that is possessed by the whole but is not by any of the constituent elements. Consciousness is an emergent property in this sense because (assuming consciousness is somehow caused or realised by physical states in the brain) none of the constituent elements of the brain display the properties of intentionality, phenomenal experience or anything else associated with consciousness. The more interesting version of this, which has genuine ontological depth, is that some property *p* of a system *x* is emergent if and only if a complete description of *x* (at any level of analysis) fails to describe or identify *p*. Consciousness is again an emergent property by this definition since a complete description of the brain at any level of analysis fails to describe or identify phenomenal experience (or so I argued in the introduction). Note that on this definition some versions of non-reductive physicalism discussed in the previous chapter (Anomalous Monism is the most obvious example) as well most versions of dualism would claim that consciousness is emergent in this sense.

Emergent complexity[2]

Before we can get to grips with the concept of emergent complexity we need to spend a little time on the notion of complexity itself. We may begin by defining a complex system as one whose parts generate a proportionately large number of possible configurations. So in the present context, for example, we might consider a person's genotype to be a relatively simple system since, assuming we ignore environmental influences, despite its inherent complexity it can generate only one phenotype. A game of noughts and crosses, in contrast, which has a relatively small number of parts (in comparison with a person's genotype), has a potential for over 50,000 legal configurations. This highlights a second feature of complex systems, that they are observer relative. If one shifts one's perspective to that of a molecular biologist, a person's genotype is part of a complex system, a human gene pool, which has billions of possible configurations. Thus complexity only becomes an issue when we try, in whatever way, to understand or model reality. There are several features of model building worth considering here, firstly a model has to be simpler than the system being modelled. By getting rid of what one hopes will be superfluous details the model builder hopes to gain some understanding of the rules and laws that govern the system. The crucial stage in the construction of any model is deciding which 'building blocks' one should use. A building block is an abstract and simplified description of an element within a system (a gene, for example, is defined as a functional unit of

DNA). If one has been successful in eliminating the superfluous detail the building block(s) should behave (in the context of the model) as the element it is modelled on behaves in the real world. In a computer simulation of genetic change in a population, for example, the genes in the model behave in a similar fashion to strands of DNA in the real population. To take an example pertinent to our present inquiry, Goffman's dramaturgical analysis is a model of social interaction and roles are its building blocks. Goffman's aim when developing the concept of roles was presumably to sheer away all the irrelevant details of persons leaving only those elements that govern social interaction. Similarly, Marx's model was dialectical materialism and his central building block, class. Successful science (including social science) depends on ensuring that one's building blocks mirror the causally efficacious aspects of the system being modelled. In part this thesis is concerned with one of the building blocks of the social sciences, the human mind. At times it may appear as though the thesis (and part I in particular) is of little relevance to the social sciences. This apparent lack of relevance derives from the fact that I am not concerned with developing a model of society, social change, social interaction or any other sort of 'social scientific' model. Rather, I am concerned with developing a model of the individual human mind. In other words, I am attempting to sheer away all the irrelevant details of persons to reveal a building block that may be of use in the construction of social scientific models.

Having considered complexity, models and building blocks, we are now in a position to return to the concept of emergent complexity. Complexity is

[2] Much of this section was influenced and informed by John H. Holland's (1998) discussion of emergence.

emergent if a small set of simple rules that govern the behaviour of the building blocks of a system are sufficient to generate complexity. Here the term emergent refers to the fact that in models there is a very real sense in which complexity is generated by the laws that govern the system. This is nicely illustrated by Holland's retelling of Hofstadter's (1979) ant colony metaphor:

> Individual ants are remarkably automatic (reflex driven). Most of their behaviour can be described in terms of the invocation of one or more of about a dozen rules of the form "grasp object with mandibles," "follow a pheromone trail (scents that encode 'this way to food,' 'this way to combat,' and so on) in the direction of an increasing (decreasing) gradient," "test any moving object for 'colony member' scent," and so on... This repertoire, though small, is continually invoked as the ant moves through its changing environment...
>
> The activity of the ant colony is totally defined by the activities and interactions of its constituent ants. Yet the colony exhibits a flexibility that goes far beyond the capabilities of its individual constituents. It is aware of and reacts to food, enemies, floods, and many other phenomena, over a large area; it reaches out over long distances to modify its surroundings in ways that benefit the colony; and it has a life-span orders of magnitude longer than that of its constituents... To understand the ant, we must understand how this persistent, adaptive organisation emerges from the interactions of its numerous constituents. (Holland 1998: 81-2)

Whether society displays this type of emergent complexity is a question that I do not intend to discuss. This question is for model builders (particularly those concerned with the structure-agency debate) to ponder. It is worth noting, however, that although no-one expects that social life will be reducible to the operation of as few as a dozen rules, there are several schools of sociological thought, both past and present, that implicitly buy into this model of emergent complexity. We have already mentioned Goffmanesque dramaturgical analysis and Marxism, but to this list we may also add functionalism, and the exemplar of this approach, rational choice theory. For rational choice theorists the enormous complexity generated by the interaction of human beings can be reduced to the laws of instrumental rationality. There is, therefore, a sense in which according to rational choice theory man displays little more flexibility than the reflex driven ant. All our behaviour can be described and predicted by rules of the form: 'never invest more (in time, energy, money or emotional commitment) than one believes one will receive in return,' 'always act in a manner that maximises pleasure and minimises pain,' and 'never pay more for a commodity (where the idea of a commodity includes such things as the love of a spouse) than is necessary.'

Despite having said that I do not intent to become embroiled in the structure-agency debate by discussing this type of emergent complexity, I should note that my discussion of the individual human mind (especially in part II) will have some direct implications for this debate. My discussion of rationality, the self and agency in part II, for example, though not dealing explicitly with rational choice theory, presents a model of the agent that is incompatible with the rational actor model.

Emergent patterns

Emergent patterns are essentially heuristic devices that allow one to predict the future state of a system based solely on recurrent patterns. Thus, one can predict the future state of a system without necessarily having any knowledge about the generating mechanisms (i.e. the causal mechanisms that generate the recurrent pattern). Emergent patterns are familiar to both the natural and social sciences as well as their folk equivalents (we are all adept at predicting the course that the common cold will follow without having the slightest medical knowledge). The economic cycles of boom and bust are a familiar example within the social sciences. A more contentious example, which we will explore in the next chapter, concerns folk psychology. If eliminative materialists are right that the categories of folk psychology (such as propositional attitudes) fail to refer to any causally efficacious physical states then it may be that our ability to predict people's behaviour derives from our having noticed certain commonly occurring patterns.

Emergent powers

It is the claim that the mind has emergent powers that threatens the version of epiphenomenalism being developed here and it is this version of emergence which we shall spend the remainder of this chapter arguing against. Before beginning our critique, it might be helpful to sketch out exactly what differentiates the theory of emergent powers from supervenience theories of mind. As I argued in chapter 2, those tempted by supervenience face a tricky balancing act between epiphenomenalism and dualism. To say that the mental supervenes on the physical is to say that mental states are *dependent* and *determined* by the physical states upon which they supervene. These dual characteristics, of dependency and determinacy, have proved (or so I argued) to be incompatible with the orthodox mental realists' claim that the mind is both *sui generis* real (i.e. irreducible) and causally efficacious. As soon as one relaxes either the dependency or the determinacy relationship, the scales tip towards dualism. On the other hand, if one retains a commitment to both dependency and determinacy then the scales become weighted in favour of epiphenomenalism.

Proponents of emergent powers (henceforth, since the remainder of this chapter deals exclusively with emergent powers, I will drop the suffix powers), in contrast, have sought to escape this quandary by relaxing the requirement that mental states be both dependent on and determined by physical states. Although emergentists are committed to the view that mental states are (at least partly) dependent on physical states, the dependency relation is typically restricted to the generation of mental states and does not, therefore, entail the property covariance characteristic of supervenience. This, in turn, gives emergentists the 'slack' they need to account for mental causation and true libertarian free will.[3] The basic idea is that consciousness emerges from physical events in the brain and thereafter enjoys a degree of autonomy such that it is able to exercise 'downward causation'. A notable exception to this is John Searle's account of system causation. Searle's view is that unless one accepts that the totality of the features of the brain at the microlevel are sufficient to fix the conscious state at that point, one would be forced to accept some form of dualism. Searle's own theory will be discussed later in this chapter.

The emergentists' rejection of the supervenience thesis makes emergence a difficult position to evaluate from a physicalist perspective. Our critique of supervenience was based on demonstrating that all the accounts of supervenient causation violate one or other of the three principles we outlined in the introduction: (P1) the irreducibility of phenomenal states, (P2) causal closure, and (P3) the principle of causal explanatory exclusion. Though emergentists tend to be advocates of (P1), and accept that overdetermination (P3) is unacceptable, they are typically at pains to demonstrate that the physical world is open.[4] Thus, if one seeks to go beyond the dogmatic assertion of physicalist principles, one is forced into providing an immanent critique. Emergentism, however, is a broad church and there are those like Searle who seek to retain a commitment to materialism and whose theories can be approached in a similar manner to those in the previous chapter.

[3] That is to say an account of free will that is based on agent causation rather than the compatibilism of authors such as Dennett (1984) and Honderich (1988).

[4] See, for example, Popper (1982) and Haskar (1999).

Before discussing the emergentists' approach to mental causation it is first worth looking at the concept of downward causation. Roger Sperry famously introduced the analogy of a rolling wheel to demonstrate how features of a system can causally determine the behaviour of its parts (Sperry 1980; Sperry 1991). The idea is that the constituent elements of the wheel, its atoms and molecules, are carried through space and time on a trajectory that is entirely determined by the properties of the system. However, as J. J. C. Smart has noted, 'to say that the motion of the particles is determined by the system as a whole is merely to say that the motion of each particle is determined by the resultant of the forces on it. If this is emergence, then this is a sort of emergence that the most reductionist and mechanist physicalist will never have dreamed of denying' (Smart 1981: 111). Another example comes from Kauffman who asks us to imagine a wrong 'decision' causing the death of Tomasina (Kauffman 2000: 129). Tomasina is the last member of a molecular species and her death robs the biosphere of the unique proteins and molecules that constituted the now extinct species. Thereafter, not only are these proteins and molecules absent from the biosphere, but so too are their descendant mutations which might have given rise to new species and novel chemical reactions. Kauffman's point is straight forward, the behaviour of autonomous agents[5] can have potential consequences for lower levels of organisation. From this, Kauffman concludes, 'Downward causation is real and nonmystical'. What does this mean? Well, it is real in the sense that an event described at one level of organisation occurred and it had consequences for future events described at a lower level of organisation. It is

[5] The definition of an autonomous agent, which is outlined in the quotes that follow, is quite technical (see Kauffman 2000: 49-109), for our present purposes a common sense understanding will suffice.

nonmystical because the downward causation exists only relative to the biologists' conceptual scheme. To explicate, Kauffman says he distrusts the bottom up version of causation that is typical of reductionism and claims that there is a sense in which whole organisms are 'more than the sum of their parts' and 'part of the furniture of the universe'. Such statements, from a respected scientist, look like lending support to those who believe that there are emergent properties capable of exercising downward causation. Once again, however, as the following quotes illustrate, the language is misleading.

> In what sense is Tomasina nothing but the atoms and their locations and motions in three-dimensional space of which she is comprised? The concepts of atoms in motion in three-dimensional space do not appear to entail the concepts of an autonomous agent, self-consistent constraint construction, release of energy, propagating work tasks... What, after all, do Newton's laws of motion have to do with a sufficient account of Tomasina's jump to the left rather than the right?' (Kauffman 2000: 129)

But contrast this with Kauffman's general account of emergence:

> ...the autonomous agent is, more than the sum of its parts, but not in the sense that the behavior of the autonomous agent is not explicable as the total organization of the parts organized into the whole agent in its environment. Rather, an autonomous agent is more than the sum of its parts in the sense that a wide variety – indeed, an infinite variety – of physical systems could be autonomous agents in the same sense, self-reproducing systems capable of carrying out at least one work cycle. (Kauffman 2000: 128-9)

The first quote relates to the possibility of prediction, and notes that Newtonian physics, or any other physics for that matter, tells us next to nothing about the creatures that inhabit (or could inhabit) our biosphere. It may also be taken to imply that the laws of biology are irreducible to those of physics. This may well be the case but it is largely irrelevant to our present inquiry. This version of reductionism is a matter for philosophers of science to ponder and has no implications for our current project. The type of reductionism and bottom up causation that concerns us here relates to singular causal statements. Any singular causal statement in biology, for example Tomasina's wrong decision, must be reducible (in principle) to causal statements that refer only to atoms in three-dimensional space. As the second quote illustrates, this type of reductionism is unambiguously accepted by Kauffman. This case nicely illustrates the danger of confusing methodological and ontological arguments. Kauffman's discussion of reductionism occurs in the context of developing an account of autonomous agents as an appropriate level of investigation and a useful conceptual tool for biology. All too often such methodological arguments are misinterpreted as making ontological claims about the reality of emergent powers.

Clearly then, the opponent of emergence need not deny that a system's features causally determine the behaviour of its parts. Emergence, if it is to differentiate itself from physicalism and make space for libertarian free will, must specify the causal relationship between the system and its parts.

William Haskar proposes the following definition:

> ... if (for example) consciousness is emergent... the behaviour of the physical components of the brain (neurons, and substructures within neurons) will be *different*, in virtue of the causal influence of consciousness, than it would be without this property; the ordinary causal laws that govern the operations of such structures apart from the effects of consciousness will no longer suffice. (Haskar 1999: 174)

The introduction of the idea of causal laws is the key step in Haskar's definition. As far as we know the component elements of the brain behave (so as not to beg the question we should prefix this with 'in the absence of consciousness') in an entirely mechanistic manner. Though some have argued that quantum events can have a causal influence on the behaviour of neurons, thus yielding a degree of unpredictability, such quantum effects have no implications for the version of emergence developed by Haskar. Quantum effects aside, therefore, it is generally assumed that one can study the behaviour of neurons and, in principle, predict their behaviour using classical (i.e. Newtonian) physics. As long as one has certain information about the neuron and its environment (the presence of neurotransmitters, the quantity and ratio of sodium and potassium ions across the membrane wall, etc.) and one knows the input, one can theoretically predict the output. If it turns out that the brain is just an aggregate of similarly mechanistic units, then there would seem to be no (in principle) reason to suppose that one should not be able to predict the brain's output based on information regarding the behaviour of each unit (again using classical physics).[6] Haskar's criterion for emergence, which is admirable for biting the bullet and stating what is certainly an unfashionable position, is that consciousness must interfere with or determine the behaviour of the constituent elements of the brain. In effect, the presence of consciousness must *violate* the mechanistic laws that would otherwise govern the behaviour of neurons. This must be the case if consciousness is to have any causal influence on the body (and not just on other conscious states), since even emergentists accept that the brain is the vehicle through which the mind interacts with the body. Note that this is a stronger claim than that the presence of consciousness has a causal influence on the behaviour of neurons. Such a claim would be accepted by all those who reject epiphenomenalism. Anyone persuaded by the identity thesis, for example, would claim that the presence of consciousness has a causal influence on neurons because consciousness is nothing but neurological events/states. There is no violation of the causal laws

[6] It is of course highly unlikely that this will ever become an empirical possibility. For such a predictive task to be successful we would not only need a vast amount of computing power but one would require a model of the subject's brain that was accurate down to the atomic level. Neurologists who attempted such a predictive task would not be afforded the luxury of approximation. While the meteorologist can afford to ignore the infamous butterfly's wings, the neuroscientist's task might be thwarted by a sodium ion mapped onto the wrong side of a membrane wall. The human brain is an astonishingly sensitive piece of equipment where the slightest difference in the initial conditions can have potentially massive short term effects (recall, for example, the case of Proust, who was prompted to write *A la Recherche du Temps Perdu* by the taste of madeleine cake).

that govern the behaviour of the component elements of the brain because consciousness is identical with the operations of some of those components (as we shall see this essentially Searle's view). Take away consciousness (by general anaesthesia or inducing a coma) and you have *ipso facto* causally influenced the component elements of the brain. Haskar's claim is stronger than this because he denies any form of identity or supervenience. For Haskar consciousness is something extra, more than the sum of its parts, and irreducible to neurological states/events at any level of description. Thus, for consciousness to have a causal influence on the brain it has to do so from *outside the system*. Haskar's own version of emergence, which he dubs emergent dualism, is to my knowledge the only well worked out theory of emergence that explicitly adopts this criterion, and it is to this that we will now turn.

Emergent dualism

Haskar adopts a realist perspective with which I have considerable sympathy. Though he rejects the metaphysical premise that the physical is causally closed, he is willing to accept the empirical findings of contemporary science (and neuroscience in particular). That is to say, in the absence of what he terms an emergent self, he takes the constituent elements of the brain to be entirely deterministic and in principle predictable (hence his insistence that, in order to guarantee true libertarian free will, the presence of consciousness must violate the laws that govern the behaviour of the constituent elements of the brain). Like myself, Haskar is also a realist about the reality of our experience of mental states and rejects any from of reductionism (including the supervenience thesis which, like myself, he takes to be incompatible with mental causation). Where my own approach and that of Haskar diverge is that Haskar is firmly committed to the principle of mental causation and free will. To summarise, Haskar proceeds from three principles: the irreducibility of mental states; a belief that in the absence of consciousness the components of the brain behave in a deterministic manner; and a commitment to the existence of free will. The latter of these three principles Haskar takes to imply agent causation. His account of emergent dualism, and the related concept of an emergent self, is an attempt to construct a theory that is compatible with these three principles.

Theories that posit some form of dualist interactionism are necessarily sparse on the ontological details of the non-physical properties and how they interact with physical properties. This is a methodological consequence of the transcendental realism which is typically the method of choice for dualists. For transcendental realists the existence of non-physical properties and powers are known only through their effects on the physical world. What we have in Haskar's account, though, is a rare and laudable description of the causal relations that obtain between the emergent self and the brain. The emergent self is presented as an entity that, though generated by events in the brain, is subsequently distinct from the events/states that brought it into existence. Haskar likens this emergent self to a *soul-field*, which, like a magnetic field, is distinct from the object that generates it. Once generated the causal relations are as follows: brain-brain interaction is deterministic, mind-mind interaction indeterministic, mind-brain and brain-mind interaction deterministic with the caveat that the effects of the brain on the mind are influenced by the mind's internal evolution (and vice versa). This allows Haskar to claim that: '... one can't say in general that a given brain input produces a certain mental effect (ibid. 200).' This gives us a picture of mind-brain interaction that looks something like this:

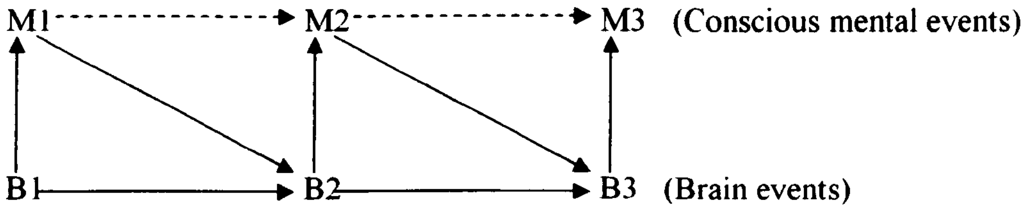


Fig 3.1 Haskar's account of mind-brain interaction

Here the arrows represent the direction of causation. Dashes represent an indeterministic relationship and unbroken arrows a deterministic relationship. (But again with the caveat that the brain-mind/mind-brain causation is influenced by the mind and brain's internal development.)

Since emergent dualism is a consistent position, given its premises, and since I have no intention of rehearsing the well known dualism versus materialism debate, I want to criticise Haskar's account by showing that he falls foul of one of his own criticisms of materialism (a criticism which, incidentally, he views as 'devastating').

Haskar correctly notes that physicalists are committed to the principle of the causal closure of the physical domain and argues that this causal closure is incompatible with the materialists' claim that conscious awareness confers an adaptive advantage. Assuming conscious states are identical with or supervenient on physical states then it follows that the causal powers of mental events are wholly determined by the physical states upon which they supervene.[7] If this is indeed the case then, Haskar argues: '*The mental properties of the event are irrelevant to its causal powers*' (ibid. 78). Haskar goes on to note that:

[7] In support of this position Haskar quotes Kim approvingly (if rather selectively).

> What this means is that, given the physicalist assumption, *the occurrence and content of conscious mental states such as belief and desire are irrelevant to behavior and are not subject to selection pressures.* On this assumption, *natural selection gives us no reason to assume that the experiential content of mental states corresponds in any way whatever to objective reality.* And since on the physicalist scenario Darwinian epistemology is the only available explanation for the reliability of our epistemic faculties, the conclusion to be drawn is that physicalism not only *has not given* any explanation for such reliability, but *it is in principle unable to give* any such explanation. And that, it seems to me, is about as devastating an objection to physicalism as anyone could hope to find. (ibid. 79)

There are a number of elements run together in this critique and it is worth spending some time teasing them out and expanding on them. Any theory which explicitly denies (such as my own) the causal powers of mental states, or entails that mental states are causally inefficacious (which I have argued is the case with the supervenience thesis), cannot appeal to natural selection in order to explain the *occurrence of any mental state or the correspondence between those states and objective reality*. In terms of correspondence there are three distinct features that require explanation:

1. The correspondence between the PE caused by the physical mechanisms that subserve perception and the object being perceived (we'll call this percept-object correspondence). For example, any theory of the mind needs to be able to explain the correspondence between my phenomenal experience of a tree (percept) and the objective features of that tree (object), its colour, height, shape, etc. In other words, given that correspondence (or lack of it) makes no causal difference, why does my visual image of a tree look like a tree and not, say, a table lamp.[8]
2. The correspondence between the PE of meaning (be that symbolic or linguistic meaning) and intersubjectively agreed meaning. In Popper's

[8] Another problematic area related to percept-object correspondence is inverted qualia. This problem, however, does not appear to have any Darwinian solution and will not be discussed here.

terminology (which for brevity we will adopt) we may characterise this as the correspondence between World 2 and World 3. In the present context we may characterise World 2 as the content of PE where that content relates to the meaningful objects that comprise World 3. World 3 consists of the products of the human mind, mathematics and mathematical symbols, language, the symbolic meaning attached to objects such as money, signs, religious imagery, etc.[9] An example of why this correspondence constitutes a problem might be, why is it the case, given that my PE is epiphenomenal, that my PE of the meaning of the term 'philosophy' (World 2 object) relates to the intersubjectively agreed meaning of the term (World 3 object).

3. The correspondence between, what I will term, the function and feel of PE. The orthodox mental realist (henceforth OMRist) view is that we evolved to experience sensations and emotions such as lust and pain (in the appropriate circumstances) because those sensations and emotions cause us to act in a way that increases our chances for survival and the chances that our genes will be passed onto the next generation. Thus, I am caused to seek out a mate because I experience lust, and I am caused to withdraw from harmful stimuli because I experience pain. The OMRist perspective is consistent with these emotions and sensations being only contingently related to their function. It is entirely consistent with the OMRist perspective and evolutionary theory, that there are a potentially infinite number of sensations that could have fulfilled the function of causing us to withdraw from harmful stimuli (and perhaps do in other species, terrestrial or extraterrestrial). Martians and monkeys need

[9] This is a more restrictive characterisation of World 2 and World 3 than Popper's formulation and by their use I do not mean to imply that I buy into Popper's ontology. I adopt them only because they are convenient and familiar locutions that are suited to the present inquiry.

not experience pain as we do. However, what evolutionary biology and the OMRist camp require is that all the potential sensations that could fulfil the function of causing us to withdraw from harmful stimuli are *unpleasant*. That is to say, a Martian would not be caused to withdraw their hand from a fire if their hand being burnt caused them to experience amusement, but they would withdraw if they experienced any sensation as unpleasant as pain. If epiphenomenalism is true the qualitative character of sensory and emotional experiences (as well as their contingent manifestation in human beings) is irrelevant to behaviour. We would behave in exactly the same manner if we experienced lust in response to harmful stimuli and pain in response to meeting a potential mate.

The reliability of our epistemic faculties is a slightly different and more complex issue. Emergentists, and dualists generally, tend to appeal to the principles of reason and rationality rather than Darwinian evolution in order to account for the reliability of our epistemic faculties. It should be noted, however, that even if one grants that such an appeal is legitimate, even the most rational mind would be rendered impotent if the correspondences outlined above were not reliable. The mind would quite simply have nothing to reason about. For the moment we will concentrate on correspondence and return to this issue in part II.

In all these cases, so Haskar would argue, mental states evolved, and correspond to objective reality, because they are causally effective and increase the organism's chances for survival and reproduction. The next question we must ask is how do these states discharge their function? For reductionists this is an easy question. Reductionists would claim that genes code for neural structures that cause us to (1) experience a correspondence between percepts and objects,

(2) experience a correspondence between World 2 and World 3, and (3) cause us to experience emotions and sensations that, in turn, cause us to behave in a manner that increases our chances for survival and reproduction. Moreover, reductionists would argue that the PE is able to discharge its function because it is (token-token) identical to the functioning of the neural structures that are coded for by specific genes. None of this is at all controversial and Haskar readily admits that if conscious states were identical to physical states their causal efficacy would be guaranteed (ibid. 74).

What is problematic is Haskar's attempt to apply the same evolutionary logic to explain the correspondences cited in 1-3. I hope to show that Haskar's insistence that the mind is emergent precludes his adopting this evolutionary argument, and this is about as much of a devastating objection to emergence as anyone could hope to find! To see why this is the case we need to return to how Haskar views the relationship between mental and physical states. In what follows we will follow through the example of the correspondence between percept and object. The arguments outlined below can be applied *mutatis mutandis* to the correspondence between World 2 and World 3 and the correspondence between the function and feel of mental states.

Despite Haskar's appeal to Darwinian evolution he does not offer us any account of the relationship between genetics and the mind or the emergent self. It is, therefore, left up to us to construct such an account. The following example concerns how emergent dualism would explain this relationship in the context of a conscious reaction to a percept. Suppose, for example, I were to step out of the road to avoid an oncoming bus, emergent dualism would explain this action as follows: as the bus approaches light is reflected of its surface, hits my retina and sends a signal to my brain. Neural structures in my primary visual cortex cause *but are not identical with* my experiencing a percept that represents the bus. This bus-percept is an emergent property (irreducible to whatever neural events caused it) and forms part of my mind. Its influence on me (on my emergent self) is dependent upon the internal evolution of the mind. That is to say, if I am awaiting the arrival of the bus because I want to take it to work, this bus-percept will provide a reason for me (but not a causally sufficient reason) for me to walk across to the bus stop and put out my arm. If, on the other hand, I have been absent-mindedly chatting to a friend in the middle of the road, the bus-percept will cause me to move out of the way in order to avoid being hit (but again it will not provide causally sufficient conditions). The crucial point to note is that Haskar would explain the correspondence between percept and the object by appeal to evolution, '...since conscious states are causally effective, they are also subject to Darwinian selection' (ibid. 197). However, by denying any form of (token-token) identity or the strict property covariance of supervenience, Haskar has broken the link between Darwinian evolution and the mind. There is nothing to tie the bus-percept to genes via the neural mechanisms that subserve perception.

To explicate, genes could not possibly code *directly* for an emergent property. Genes could only code for neural structures that might give rise to an emergent property. But since the point at issue is not the existence of these emergent properties but to explain their correlation with objective reality, a close correlation between the emergent properties and their physical cause is required. On emergent dualism genes code for neural structures that give rise to perceptual experiences. These perceptual experiences are emergent phenomena, irreducible to the neural events that brought them into existence. Once created these emergent phenomena float free of their neural cause. If their existence benefits the organism to which they belong then it is conceivable that the genes that coded for the neural structures that caused the emergent phenomena will be passed onto the next generation. At this point it is important to note that any causal powers that the emergent phenomena possessed, their correspondence with objective reality and any evolutionary advantage they conferred on the organism in which they were instantiated are not passed on (directly) to the next generation. Rather, the neural structures that gave rise to the emergent phenomena is the hereditable trait. Now, if it turns out that the next generation does inherit these neural structures, and that these neural structures again provide an evolutionary advantage in virtue of the emergent phenomena they cause, then it looks like we have at least minimal property covariance between the emergent phenomena and their neural cause. Such property covariance is incompatible with the libertarian free will that Haskar's account is designed to secure. Property covariance of this sort would amount to a form of supervenience and this I have argued (and Haskar readily accepts) amounts to determinism, epiphenomenalism, and the subversion of free will.

Despite the above criticisms emergent dualism remains a consistent and viable option. The point of the above discussion was not to provide a refutation of emergentism at the dualist end of the spectrum (since this is ruled out by the three principles outlined in the introduction). Rather, the point of the discussion was to show that emergentists cannot occupy a halfway house between outright dualism (of the Platonic/Cartesian tradition) and materialism by appealing to Darwinian arguments.

System causation

If one accepts that conscious states are caused and realised by physical states in the brain, that is one accepts that conscious states are entirely determined by physical events in the brain, how is libertarian free will possible? For Searle libertarian free will, and we can agree with this point, is incompatible with our actions being determined by causally sufficient conditions, be they psychological (such as, for example, that the desire for a glass of wine is causally sufficient for me to drink a glass placed in front of me) or physical (hence the rejection of any form of compatibilism). True libertarian free will requires that I, as a conscious subject, am the cause of my actions and this would seem prima facie to be at odds with the claim that all my conscious states are caused and realised by physical events in the brain. Searle's intriguing solution to this problem begins by noting that during voluntary actions there are three gaps where we do not experience our psychological states as causally sufficient to determine our actions. Searle then goes on to suggest that these gaps in causally sufficient conditions exist at the neurobiological as well as the psychological level (thus escaping physical determinism and making room for libertarian free will). The final step is to postulate the existence of an irreducible self which acts on reasons but is not causally determined by them.

Following the observation that we experience the world from the perspective of a unified self, Searle notes that our experience of action contains a lack of causally sufficient conditions. 'I do not experience my reasons – my beliefs and desires, for example – as causally sufficient to fix one decision rather than another… [Nor do I] experience the decision as causally sufficient to produce the action I have decided upon' (Searle 2000: 7). Thus, there are two 'gaps' where we experience the self as acting voluntarily. There is the gap between reflecting on reasons, beliefs and desires, and making a decision and there is a gap between making a decision and acting on the decision. In the case of actions that are extended in time then there is a third gap during which we have to choose to continue performing the action. These gaps do not entail that reasons and decisions do not function causally, rather, that 'the antecedents function causally, but they do not function by way of causally sufficient conditions' (ibid.).

We will examine these gaps in detail later, before doing so it is worth reflecting on the reasons Searle presents for postulating the existence of a self. Searle's conclusion (which he admits he came to with some reluctance) is that an ineliminable and irreducible self is presupposed by the nature of explanations of behaviour. Searle compares two types of explanation; both based on reasons, but in the case of the former involving free will:

A: I made a mark opposite Jones' name because I wanted to vote for Jones.

B: I got a stomach ache because I wanted to vote for Jones.

Searle notes that there is a difference in the logical structure of A and B. Sentence B takes the form: Event (or state) *x* caused event (or state) *y*. Sentence A takes the from: A self *S* performed act *A* because *S* was acting on reason *R*. Sentence B, but not sentence A, states causally sufficient conditions, yet according to Searle both sentences are adequate explanations and the only way to explain this adequacy is to posit the existence of a self.

Searle is quite right here, that *if* sentences of the logical form A were adequate explanations for action then an irreducible self would be required to account for their adequacy. However, if one rejects the claim that reasons are causes (that sentences of the logical structure of A are adequate explanations) then one is not forced to accept the conclusion that there must be an irreducible and ineliminable self. This topic will be the focus of later chapters, for the present analysis we will accept Searle's conclusion. I should note, however, that I tend to think that the logical structure of action sentences is rather flimsy evidence for the existence of an irreducible self. It seems to me that we would be better off questioning the logical structure of action sentences rather than inventing entities and powers that make them true.

We can agree with the point (which will be taken up in chapter 9) that psychological causes are not causally sufficient for action: 'A complete specification of all the psychological causes operating on me at time t_1, with all their causal powers, including any psychological laws relevant to the case, would not be sufficient to entail that I would perform act *A* under any description' (ibid. 12). Searle's idea is that if this lack of causally sufficient conditions at the psychological level were matched by a lack of causally sufficient conditions at lower levels (e.g. neurological, molecular, quantum) then this might provide the gap necessary for the exercise of libertarian free will.

Before examining Searle's proposal there are a couple of preliminary points to make. Firstly, Searle's position is dependent on the identity thesis. Searle is committed to the idea that bottom-up neurobiological processes in the brain 'cause and realise' consciousness. The idea Searle has in mind is that psychological causes and neurobiological causes are 'the same causal sequences described at different levels' (see ibid.14-15). The phrase 'causes and realises' is a little misleading. The concept of a cause presupposes a subsequently occurring effect; thus, we would have a neurobiological cause and a psychological effect.

Realises, though, suggests some form of identity where psychological properties are realised by a synchronous neurobiological 'cause'. We can clarify matters by considering Searle's approach to levels and bottom up and top down causation. Searle claims that these terms are misleading and that:

> Consciousness is no more on the surface of the brain than liquidity is on the surface of water. Rather the idea that we are trying to express is that consciousness is a system feature. It is a feature of the whole system and is – literally – at all of the relevant places of the system in the same way that the water in a glass is liquid throughout. (Ibid. 16).

Searle's idea is that consciousness is realised by the whole system and that the gaps in causally sufficient conditions go right down to the 'bottom'. Quite why we should think of the brain as an indeterministic system is left open, but Searle suggests that if we abandon 'prejudice' of thinking of the brain in terms of neurons and go right down to the quantum mechanical level then indeterminacy does not seem so puzzling (see ibid. 17).[10] Thus, because consciousness is realised by the whole system, causal explanations that refer to conscious states also refer to the whole system (at every level). Rational agency, for example, is '...realized in neurobiological structures that have these properties *as well*, that are *themselves* the underlying structure of rational agency...'(ibid. 19 emphasis added). This mapping of conscious psychological processes onto lower level physical processes (presumably including the quantum level) is essential to prevent a lack of causally sufficient conditions logically entailing randomness. Quantum indeterminacy was introduced as a means of escaping determinism, but for Searle's version of system causation indeterminacy is not enough. Indeterminacy is a necessary but not sufficient condition for free will since on its own indeterminacy merely introduces randomness. To guarantee free will, consciousness and the self must be caused and realised by the whole system at every level (including the quantum level).

Although this is a very appealing analogy, the idea that consciousness is a system feature that exists at every point in the system just as water in a glass is liquid throughout looks like a claim about the identity of consciousness with the system. Such an identity claim brings with it a host of problems relating to reductionism. The trouble is that identity goes hand-in-hand with reductionism and Searle is emphatic that consciousness is irreducible. We have already discussed Searle's anti-reductionist arguments and found them compelling. We have also noted that Searle claims these arguments are ontological not epistemic. However, after presenting his anti-reductionist arguments, Searle, as the following quotation illustrates, goes on to note that they have no deep consequences (see Searle 1992: 118-124):

> ...once the existence of (subjective, qualitative) consciousness is granted... there is nothing strange or mysterious about its *irreducibility*. Given its existence, its irreducibility is a trivial consequence of our definitional practices. Its irreducibility has no untoward scientific consequences whatever. Furthermore, when I speak of the irreducibility of consciousness, I am speaking of its *irreducibility according to standard patterns of reduction*. No one can rule out a priori the possibility of a major intellectual revolution that would give us a new – and at present unimaginable – conception of reduction, according to which consciousness would be reducible. (Searle 1992: 124)

As I understand it, the reference to untoward scientific consequences concerns epiphenomenalism and dualism. Searle's point is that despite its irreducibility consciousness can be treated as a biological phenomenon like any other, hence irreducibility has no deep consequences. Once can discern a subtle shift from ontology to epistemology here. The fact that the irreducibility of consciousness has no deep consequences derives, on this account, from the fact that it is

[10] There is some disagreement as to whether quantum indeterminacy could have any effect at the level of neurons. For an account of the possible role of quantum events in brain function see

irreducible only according to the standard patterns of reduction, and this is surely an epistemological claim. When one is concerned with the possibility of an ontological reduction there is no middle ground – either consciousness is reducible or it isn't. Searle seems to want to have it both ways; he tells us that where consciousness is concerned the appearance is the reality. This is an ontological claim with which we have agreed and it is true or false irrespective of the scientific concept of reductionism. On the other hand, Searle's version of system causation requires that consciousness can somehow be identified with the physical processes in the brain which cause and realise it. It seems that Searle is advocating a position that may be characterised as one of numerical identity with property dualism. That is to say, consciousness is identical with the system which causes and realises it, but possesses properties irreducible to that system. There is nothing inconsistent in the claim that a single physical entity can possess both physical and mental properties, but it is inconsistent to maintain that consciousness is both identical to the system and irreducible to that system.

Before going on to outline Searle's theory of system causation it is first worth outlining what Searle views as the alternative model of the relationship between the three psychological 'gaps' and neurobiology. Since the title of this model, psychological libertarianism with neurobiological determinism, is self explanatory we can quickly present the hypothesis using a slightly modified version of Searle's diagrammatic representation of the model.

Penrose (1994), and for a critique of Penrose see Grush (1998).

$$\begin{tikzcd}
\text{deliberation on reasons} \arrow[rr, dashed, "\text{causes with gaps in sufficient conditions}"] & & \text{decision} \\
\text{neurobiological states} \arrow[u, "\text{C\&R}"'] \arrow[rr, "\text{causally sufficient for}"] & & \text{neurobiological states} \arrow[u, "\text{C\&R}"']
\end{tikzcd}$$

Fig. 3.2 Hypothesis 1: psychological libertarianism with neurobiological determinism. Previous conventions apply in this diagram: dashes represent an indeterministic relationship, solid lines a deterministic relationship and the arrows represent the direction of causation. C&R means causes and realises.

The reasons for presenting hypothesis 1 will become clear later. At this point I shall outline Searle's second hypothesis, system causation with consciousness and indeterminacy. According to this hypothesis the lack of causally sufficient conditions goes all the way down to the neurobiological states that cause and realise the conscious states. It is here that the self (a term which Searle uses reluctantly) comes into play. Unfortunately Searle does not have much to say about the neurobiology of the self. We are told that:

There is an x such that:

1. *x* is conscious.
2. *x* persists through time.
3. *x* operates with reasons, under the constraints of rationality.
4. *x* operating with reasons, is capable of deciding, intending, and carrying out actions, under the presupposition of freedom.
5. *x* is responsible for at least some of its decisions. (Searle 2001: 95)

When it comes to the neurobiology of the self we are left, more or less, to our own devices. Nevertheless there are several obvious features that are entailed by

Searle's description of the self in conjunction with the claim that the self is caused and realised by neurobiological states:

6. *x*, like all conscious phenomena, is caused and realised by neurobiological states.
7. (From 2.) the neurobiological structures that cause and realise the self persist through time.
8. (From 3.) the neurobiological structures that C&R the self act on certain neurological states that cause and realise beliefs, desires, reasons and the rest. This point is, more or less stated by Searle when he claims: 'In the end when we talk about consciousness affecting other elements, we are really just talking about how the elements affect each other because the consciousness is entirely a function of the behaviour of the elements…'(Searle 2001: 297)
9. (From 4.) the neurobiological structures that decide, intend and carry out actions, do so in the absence of causally sufficient conditions.

Point 9 yields the following problem: How does the *self*, so conceived, operate such that it is able to causally influence *other neurological events* in such a way as to be in accord with the its own desires. (This convergence of two language games, neurobiological and mental, may sound odd, but it is consistent with Searle's hypothesis and helps to avoid some long and clumsy locutions.) A similar point is made by Searle when he notes that the problem with his hypothesis is in explaining how the consciousness of a system can give it a causal efficacy without being deterministic (Searle 2001: 297). In other words, if as a neurological structure operating through time the self provides causally sufficient conditions for decisions and action (contrary to Searle's hypothesis) then the system would be deterministic and would not, therefore, allow room for libertarian free will. If the self does not provide causally sufficient conditions then what causes the system to act in accordance with the self's desires. If the system is to remain indeterministic in virtue of quantum indeterminacy then it looks like all Searle's proposal amounts to is a theory of random decision making and choice of action. Again this is something which Searle accepts is inconsistent with free will.

Returning to our troubles with Searle's anti-reductionism, there is still the problem of accounting for how conscious states can be causally efficacious given that they are irreducible to physical states in the brain. The reason that this is problematic is not because we fail to understand Searle's claim that conscious states are system features that are causally efficacious because they are caused and realised by events in the brain. Rather, the problem stems from one of Searle's reasons for rejecting hypothesis 1. Searle claims that hypothesis 1 leads to epiphenomenalism because it '...has the consequence that the incredibly elaborate, complex, sensitive, and – above all – biologically expensive system of human and animal conscious rational decision making would actually make no difference whatever to the life and survival of the organisms' (Searle 2001: 286). Given Searle's position, which we have characterised as one of numerical identity with property dualism, this does not follow. On this thesis neurobiological processes cause and realise conscious experience, our experience of the gap, rational decision making, etc. The only difference between hypothesis 1 and hypothesis 2, in relation to the causal powers of reasons, beliefs and desires, is that on hypothesis 1, but not hypothesis 2, the conscious deliberation on reasons and decisions are caused and realised by neurological states that form part of a *total* neurobiological state at T1 that is causally sufficient for the neurobiological state at T2. On the second hypothesis, the neurobiological state that causes and realises reasons has to be acted upon by an irreducible self in order to move the system (without introducing causally sufficient conditions) to the state at T2. If hypothesis 1, but not hypothesis 2, entails epiphenomenalism, this can only be because the self (in hypothesis 2) contributes something radically different in kind from the neurobiological cause and realisation of reasons in hypothesis 1. However, if the self is nothing more than a neurobiological process then it is no better off with regard to causal efficacy than reasons are in hypothesis 1.

To make this point absolutely clear, on hypothesis 2 the self is conceived as an entity which is caused and realised by the whole system (down to and including the quantum level). In hypothesis 1 all mental states are again caused and realised by the whole system (down to and including the quantum level). Searle claims that mental events in hypothesis 1 are epiphenomenal and that the self in hypothesis 2 is causally efficacious in moving the system according to its own desires (the self's desires must, of course, be realised by the whole system). What Searle has been unable to do is to provide *any* explanation for either the powers of the self in hypothesis 2, or why mental events in hypothesis 1 are epiphenomenal but the self in hypothesis 2 is causally efficacious. Without something to fill this explanatory gap Searle's hypothesis is little more than wishful thinking.

Both Haskar and Searle have provided detailed accounts of the causal relationships that would have to obtain between mental and physical events in order to guarantee free will and mental causation. Haskar appeals to emergent powers but, in common with all other dualist accounts, fails to explain either the origin of these powers or how they causally influence physical events. His appeal to Darwinian evolution, which I argued was illegitimate given the autonomy of the emergent self, only serves to illustrate the deep divisions between materialism and dualism which have only been glossed over by talk of emergent powers. Searle is no more successful in demonstrating the causal powers of an irreducible self, indeed in some ways he is even less successful than Haskar. By retaining a commitment to supervenience[11] Searle denies himself the escape route of emergent powers and without an explanation of how the self guides the system his appeal to quantum indeterminacy looks like introducing randomness while leaving free will and mental causation as elusive as ever.

The moral of the last chapter was that supervenience ends up either entailing epiphenomenalism or straying into the territory of emergence. The moral of this chapter is that emergence ends up either collapsing back into supervenience, with its attendant problems (as with Searle), or straying into the territory of dualist interactionism (as with Haskar). The conclusion is that neither non-reductive physicalism, as typified by the theories of supervenient causation discussed in the previous chapter, nor emergence offers any middle ground between reductionist physicalism and dualist interactionism. So where does this leave us and what choices do we have? Well, if you endorse supervenience and want mental causation, then it looks like you are going to have to give up premise 1 and accept that mental states are reducible. If emergence is more to your taste then you have two options. If you want emergence and mental causation then you are going to have to give up premise 2 and embrace a form of

[11] Searle is quite scathing about the concept of supervenience claiming that it does no philosophical work. However, despite not using the term himself, he does endorse the central tenets of strong supervenience (see Searle 1992: 124-6)

dualist interactionism. Alternatively you could resolutely refuse to give up any of your materialist principles, but then you will be stuck with emergence with epiphenomenalism. Either way, if the last fifty years of philosophy have taught us anything, they have taught us that we are not going to be able to have our cake and eat it. Despite all the recent advances in the philosophy of mind we still face the same set of problems that confronted the traditional versions of dualism and materialism. The only difference is that in today's concept rich world to see these problems clearly we have to cut through an awful lot of undergrowth. That is not to say that progress has not been made in recent years. We were always going to have to go down a lot of dead ends before finally solving the mind-body problem – at least now they are marked on the map.

Chapter 4

Folk Psychology, Connectionism, and Eliminative Materialism

Epiphenomenalism deals folk psychology a serious blow. Central to all concepts of folk psychology is the assumption that people behave the way that they do because they have the conscious experiences they do. Once this principle has been denied all that's left is an account of the role played by the content of mental states. Eliminativists, however, believe that the fatal blow against folk psychology has already been struck – epiphenomenalism notwithstanding. One prong of the eliminativists' attack is the claim that if the brain turns out to be a connectionist network then folk psychology is wrong about the causes of our actions and the elimination of its ontology (propositional attitudes, propositional memories, and so on) follows. In this chapter I want to argue that not only is folk psychology under no threat from connectionism but neither does epiphenomenalism present it with any insurmountable problems. In making this case we need to be absolutely clear about what is meant by folk psychology. To that end I will spend some considerable time outlining what I take to be three different versions of folk psychology: what I will term the folks' folk psychology, the philosophers' folk psychology and the computational theory of mind. We will explore these three versions in detail below, but briefly, the folks' folk psychology (henceforth the FFP) is the psychology used by ordinary folk in their day-to-day interaction with the world. The philosophers' folk psychology (henceforth PFP) is the philosophers' and psychologists' ontological elaboration of the FFP. And the computational theory of mind is a specific empirical thesis (or family of theses) which claims that at the implementation level cognitive thought consists of rule governed symbol manipulation. Of the above three versions of folk psychology (henceforth FP) only the latter empirical theory would be falsified if connectionism is true. The falsification of this empirical theory, however, cannot possibly warrant the elimination of FP's ontology. Eliminative materialism is only a plausible doctrine if one identifies FP with the computational theory of mind or, at the very least, if something akin to the computational theory of mind is entailed by FP. I will be at pains in this chapter to argue that any such identification does a disservice to ordinary folk and that most philosophical accounts of FP are more or less consistent with even the most radical versions of connectionism.

Given epiphenomenalism, if mental states are to have any explanatory role it must be because they are linked to other mental states or action by a form of epiphenomenal or supervenient causation. Although we have concluded that PE must be linked to future mental states and action by epiphenomenal causation, this causal relation is of no practical use. The conclusion that PE is epiphenomenal means that regardless of how reliably a given PE covaries with an action (such as phenomenal pain and withdrawal) this covariance cannot be exploited. It cannot be exploited, of course, because we have no access to PE. Having ruled out PE as a predictive and explanatory tool we are only left with mental content. Mental content, we have concluded, may be linked to the content of other mental states or action by supervenient causation. Thus, although the belief (defined in terms of its PE) that the car is low on fuel does not cause the driver to pull into a petrol station, the belief (defined in terms of its content) may be causally relevant if it supervenes on physical states that do cause the driver to pull in. Moreover, in order to be explanatorily useful, mental kinds must covary in a systematic and reliable way with physical kinds. So, for example, the belief that the car is low on fuel, if it is to be explanatorily useful, must covary in a systematic and reliable way with physical states that cause the driver to pull in. This property covariance, however, need not instantiate the kind of strict laws that Davidson argues against in proposing the thesis of anomalous monism.

Connectionism, however, poses a threat to even this restricted concept of mental causation. If it turns out that the brain really is a sophisticated connectionist system then some have argued that mental content, propositional memories and propositional attitudes, will find no place in a completed scientific ontology. Since many people working in this field seem willing to hand over the keys of our ontological toolkit to science, no place in a completed scientific ontology means no place in any ontology. Thus if cognitive researchers can't find beliefs anywhere in the brain, psychologists, sociologists, and the man on the Clapham omnibus will all be expected to stop using the term just as we have stopped crying 'witch' every time we suffer some misfortune. In line with what I have said earlier (following Searle) about the appearance being the reality, I think that this abrogation of responsibility is a mistake. In short, as long as the man on the Clapham omnibus believes that he has beliefs then science can not tell him otherwise.

We will return to this issue later, at this point I want to sketch an account of how mental content might be causally relevant. The first point to note is that a theory of causal relevance must begin with an account of how mental states map onto the physical states upon which they supervene and which are the putative cause of the behaviour one is seeking to explain. Thus, if one wants to explain behaviour by citing beliefs, desires, and other propositional attitudes, one must first provide an account of why those properties are explanatorily relevant.

Before doing so, however, I want to introduce a distinction between occurrent and dispositional propositional attitudes. Occurrent propositional attitudes are states that occur *now* and are experienced as such. Unless stated otherwise when I refer to such occurrent states I mean to include both their PE and their subvenient base. We will consider the nature of dispositional propositional attitudes in detail below, for now it is enough to note that when we attribute a dispositional propositional attitude to a person we are referring to a disposition to hold an occurrent propositional attitude. In most of what follows I will focus on beliefs but everything I say can be generalised to cover the other propositional attitudes and propositional memories. Causal relevance:

> The PE of believing (the *what it is like* to believe that ___) covaries in a systematic and reliable way with certain neuronal states (we have concluded that this must be the case from our discussion of supervenience). The neuronal states upon which consciously held beliefs supervene are necessary and sufficient for the behaviour that is traditionally explained by those beliefs (*ceteris paribus*[1]). Moreover, in cases where behaviour is explained by dispositional beliefs, the behaviour is caused by neuronal states that are *functionally similar* to those that provide the subvenient base for occurrent beliefs. The property covariance is such that one can predict, with a reasonable degree of accuracy, the behaviour caused by the subvenient base of dispositional beliefs based solely on our having correctly ascribed a dispositional belief (plus relevant contextual information) to a person.

If our conclusions from the last two chapters are accurate then the above account is a best case scenario. On this account, we lose the *mental* (in the form of PE) from mental causation but, as we saw in previous chapters, this is often an unstated consequence of contemporary materialist theories of mental causation. We will assess in this chapter how realistic an account of mental relevance this is.

At this point, it is probably also worth outlining the worst case scenario – mental irrelevance.

> Even on the worst case scenario occurrent beliefs covary in a systematic and reliable way with certain neuronal states (due to supervenience). However, the

[1] In what follows all such claims should be read as being hedged by *ceteris paribus* clauses.

> neuronal states upon which occurrent beliefs supervene are neither necessary nor sufficient for the behaviour they are held to cause. Moreover, when we refer to dispositional beliefs rather than occurrent beliefs, these terms fail to refer to *any* physical states. Beliefs, therefore, are irrelevant to the prediction and explanation of behaviour.

If this scenario, or some variation of it, turned out to be true, then it seems inevitable that propositional attitudes would go the same way as luminiferous ether, phlogiston, and witches. Propositional attitudes, in short, would be eliminated from our scientific ontology. This would be a *worst* case scenario, of course, because it would wipe out much of the subject matter of the social sciences in a single stroke. My hunch is that the truth will lie somewhere between these two extremes. Though I will argue that propositional attitudes are here to stay, I will suggest that the class of events they are invoked to explain will shrink dramatically. This, as the eliminativists are fond of reminding us, is an ongoing trend. We no longer, for example, have recourse to intentional states in order to explain the movements of the planets or the aetiology of disease and natural disasters. I will argue, however, that it is unlikely that propositional attitude terms will be eliminated from any of the disciplines in which they are currently employed.

In the first two chapters we adopted mental realism as a working hypothesis. That is to say, although we questioned the causal efficacy of PE, we assumed the *existence* of mental states as they have traditionally been understood. We have traditionally assumed the existence of three broad and interrelated categories of mental states: perceptual states derived from both proprioception and exteroception (pain, smell, touch, sight, etc.); emotional states; and cognitive states such as propositional attitudes. Each of these classes, according to common sense psychology, has a causal influence on our future mental states and behaviour. With the burgeoning of interest in connectionism in the early eighties the very existence of these mental states, and the status of common sense psychology, was put in doubt. Interestingly, attention has focused almost exclusively on the latter class mentioned above but connectionism also poses questions for our common sense understanding of the causal role of perceptual and emotional states. However, unlike propositional attitudes there is no well worked out theory regarding the casual powers of these classes and since, as we shall see, they are probably not candidates for elimination anyway, attention has focussed elsewhere. It is worth highlighting the link between theory and elimination at this point. The standard justification for elimination is to show that a theory in which certain properties or entities figured prominently is mistaken. The rejection of the theory then entails the elimination of the properties or entities which figured in the theory unless a place can be found for them in a successor theory. The trouble is that although theories abound in the cognitive sciences in which propositional attitudes and propositional memories figure prominently, ordinary folk do not hold such theories. Or at least they do not hold the type of theory whose rejection would entail the eliminativists' conclusions.

Despite nearly twenty years of debate there is still very little consensus regarding the implications of connectionism for FP. This lack of consensus is due partly to the wide variety of folk psychologies on offer. There is little agreement about the scope or ontological depth of FP, whether it is a set of rough and ready generalisations or a causal-explanatory model, for example. More importantly though, there is disagreement about the properties and powers that physical states would require in order for them to count as the referents of the posits of FP (i.e. the referents of beliefs, desires, and the rest). Specifically, there is some debate as to whether FP is committed to propositional modularity – the view that propositional attitudes are functionally discrete and semantically interpretable states. To phrase this in terms of supervenience, there is some debate as to whether propositional attitudes supervene on neurophysiological states that are semantically interpretable and causally active on some occasions and causally inactive on others. This latter feature means only that, for example, my belief that water boils at 100°C has no causal influence on my decision about when to mow the lawn. The only area of agreement seems to be in setting the minimal requirements for FP to be considered as a candidate for elimination. Viz. FP must be a theory that conceives of cognitive thought as rule governed symbol manipulation. Our first task, when examining the implications of connectionism for FP, therefore, will be to develop an account of FP and test whether these requirements are met. We well also need to be sensitive to the theory of reference that is being employed by the various commentators on FP. Steven Stich (1996) has persuasively argued that before one can come to any conclusion regarding the elimination or retention of folk psychological categories, one owes the reader a theory of reference. Unfortunately most commentators are not careful to provide an explicit theory of reference. This problem is particularly acute when discussing ordinary folks' (as opposed to academics') conception of common sense psychology.

Though the debate surrounding connectionism may seem far removed from the social sciences it does have some direct and fundamental implications that are worth keeping in mind. Not only is part of the ontology of the social sciences on trial here but the very status of language within the social sciences is called into question. We have already accepted that all human behaviour is caused by, and in principle explainable solely in terms of, the behaviour of the most fundamental particles and their properties. Accepting this microphysical determinism does not, however, preclude the possibility that human thought and behaviour might be determined by what some view as language dependent phenomena such as concepts, beliefs, and social rules. If beliefs, desires and the other propositional attitudes do supervene on functionally discrete states that are semantically interpretable, then one can still talk of beliefs superveniently causing actions. Moreover, if our acquisition of beliefs is somehow dependent on (again I am talking about supervenient causation here) language then one can still embrace epiphenomenalism while espousing the 'culturist' traditions of, for example, Winch, hermeneutics, or structuralism. To elaborate, no-one disputes that learning a language or acquiring a belief involves physical changes in the brain. If it turns out that these changes create, or create the potential for the instantiation of, functionally discrete states that interact in the way that they do because of the linguistic world that the subject inhabits, then not only is consciousness irrelevant to the study of man and society but so too is man's biological 'hardware'. This is not a debate that I intend to enter into. I merely wish to note that the possible truth of hermeneutics, structuralism, Winchian social psychology, rational choice theory, and a good many other 'isms' is dependent on propositional modularity. If it turns out that learning a language, acquiring beliefs and so on, does not create physical states that display propositional modularity (as connectionists maintain) and hence are the 'language of thought', as Wittgenstein would say, then language and beliefs are merely the vehicle for the expression of thought that exists independently of those phenomena. This latter scenario would allow man a great deal more autonomy over culture and society since, put simply, his thoughts would be his own rather than being a 'gift of society'. Very crudely, the computational theory of mind, and to a lesser extent the PFP, supports the former hypothesis and the connectionist model the latter.

Connectionism: a brief introduction

The nature of the debate between the connectionist motivated eliminativists[2] and the defenders of the computational theory of mind and FP makes it almost impossible to provide a description of connectionism that does not beg the question. I therefore caution readers that are unfamiliar with the subject not to make assumptions based on the following description and to take note of the liberal sprinkling of scare quotes.

Although connectionism has been around for some time, it was only in the early eighties that it emerged as a serious challenge to the dominant paradigm within the philosophy of mind and cognitive psychology, viz. the computational theory of mind. The computational model of cognition (which for brevity we will call the classic model) is based on the idea that the mind operates in the same way as a digital computer with cognitive processes being conceived of as rule governed symbol manipulation. To explicate, on the classic model cognitive thought consists of the syntactically driven interaction between neurally realised representations. It is for this reason that propositional modularity is such a corner stone of the classic model for it is the structural properties of the physical instantiations of representations (in conjunction with syntactically realised rules) that causes the system's behaviour (see Fodor 1988).

[2] Connectionist motivated eliminativists, like the Churchlands, argue that if connectionism is true propositional attitudes will have no place in a completed *scientific* ontology and should, therefore, be eliminated from *all* ontologies.

It was hoped that the study of the rules that enable the simulation of cognitive processes in computational models might shed some light on the rules that govern the same cognitive processes in humans. After forty years of research based on the classic model though, researchers faced what has been described as Kuhnian crisis. The classic model proved to be too rigid and deterministic to account for the flexibility of human cognitive processes and was seemingly incapable of performing many of the most basic human abilities (such as, for example, pattern recognition). These shortcomings led researchers to turn their attention to connectionist networks (or neural network/parallel distributed processing models, as they are sometimes called).

Connectionists abandoned the idea that cognitive processes are governed by explicit rules in favour of an approach that viewed 'high level' cognitive processes as arising from the interaction between simple 'neuron like units'. Connectionist networks are built from layers of units that are 'functionally similar to neurons'. However, as Berkeley has warned us, there are important differences between the units in a connectionist network and neurons which makes the claim that connectionism is a biologically plausible model questionable. Briefly, in a connectionist network each unit is connected to *all* the units in the previous and subsequent layers. Neurons, in contrast, are typically only connected to around three per cent of the other neurons in the surrounding 1mm of tissue. This makes connectionist systems massively parallel in comparison to the relative sparsity of connections in a neuronal network (see Berkeley 1997). Moreover, as Berkeley notes, there are twelve different types of neuron in the neo-cortex alone which each have different firing patterns. Some neurons have oscillatory firing patterns, some have firing patterns that are a function of their recent history and some spike randomly (i.e. in the absence of any input). None of these subtleties, and Berkeley cites many more, are mirrored in connectionist networks whose connections are in the form of continuous numerical values.

One of the main advantages of connectionist models over symbol manipulating models is that connectionist networks are capable of 'learning'. One of the most commonly used methods enabling connectionist networks to learn is back propagation. This method uses a learning algorithm during 'training' to adjust the connection weights of the units. Initially a network is presented with and processes an input. The output of the system is then compared to the desired output and an error signal is derived. The error signal is then fed back through the system adjusting the connection weights in the direction of the desired outcome according to the learning algorithm. Though there are some parallels with Hebbian strengthening and feedback connections,[3] many commentators have concerns about the biological plausibility of these learning procedures and note that it looks more like programming than learning. These concerns seem warranted when we consider that training a network to perform even simple tasks can involve hundreds of rounds of training. It seems fair to say that so far connectionist systems are far from being realistic models of neural networks.

Nevertheless, despite these shortcomings, connectionism has generated tremendous excitement within the fields of cognitive psychology and the

[3] Hebbian strengthening is the tendency of neurons within a network to strengthen their connections after a period of excitation by increasing their stockpile of neurotransmitters or growing extra synapses. This then enables the neurons to reform the network and fire together again in the future (see Hebb 1949). Feedback connections are loops of nerve fibres that connect cells receiving input with the original firing cell.

philosophy of mind. The features of connectionism that are of interest to philosophers are thankfully straight forward and do not require any familiarity with the applications to which connectionist systems have been put or the structural details of connectionist networks. For our purposes we need only note that representations/propositions in connectionist networks, unlike the classic model, are widely distributed and not semantically interpretable, or so it is claimed (we will examine this claim later in the chapter). The easiest way to appreciate these features is by comparing the way that information is represented in classic and connectionist models. In classic models, each representation, and each element in a proposition, is functionally discrete. Problem solving consists of manipulating these functionally discrete elements according to explicit rules. Moreover, in classic models information is stored in such a way that it is semantically interpretable. Crudely speaking, this means that someone with the appropriate technical knowledge could identify a representation in a classic system by citing the spatial co-ordinate(s) of the representation – they could say, 'that *means* ______.' This does not entail that a representation in a classic system must be stored at a single point. Rather, it entails that information may be accessed independently of other information and that a single item of information (e.g. a belief) may be removed without affecting the rest of the system (see van Gelder 1991). Connectionist systems, in contrast, store information holistically in the form of, what van Gelder terms, superposed schemes. In a superposed scheme, all the information is stored by the whole system without it being possible to make any fine-grained distinctions. 'Thus in connectionist networks we can have different items stored as patterns of activity over the same set of units, or multiple different associations encoded in one set of weights' (van

Gelder 1991: 43). The upshot of all this is that if connectionism is true all our beliefs, desires, propositional memories, and so on, are causally implicated in all our actions. So, for example, my belief that water boils at 100°C is just as casually efficacious in my decision to mow the lawn as my belief that the lawn needs mowing. This, the eliminativists claim, means that connectionist systems cannot exhibit propositional modularity and FP must, therefore, be false.

Some have objected that propositional attitudes might be encoded in connectionist systems as activation patterns and that activation patterns might count as functionally discrete and semantically interpretable states. Moreover, that occurrent beliefs are realised by synchronous activation patterns, and dispositional beliefs are realised by the systems' disposition to go into a state that is a realisation of the belief in question. There could be some truth in this assertion which would dovetail nicely with both supervenience and the thesis of multiple realisation considered earlier. We will return to this issue towards the end of the chapter but at this point I want to consider the eliminativists' reply. Ramsey, Stich and Garon object to this possibility on the grounds that activation patterns are not enduring states of the system. People hold, they claim, an enormous number of beliefs over a long period of time but activation patterns are transient states and cannot, therefore, be identified with the instantiation of beliefs in a connectionist system. Moreover, while it may be true that connectionist networks have an enduring 'disposition' to realise a particular activation pattern, this disposition cannot be a physical realisation of dispositional beliefs. The reason, they claim, is that they are not the 'right sort of enduring states – they are not the discrete, independently causally active states that folk psychology requires' (Ramsey et al. 1996: 111). But by whose standards are they not the right kind of enduring states? Certainly not, I shall argue below, by the standards of ordinary folk. Since ordinary folk do not claim to know what the referents of beliefs are, activation patterns and dispositional activation patterns would suit ordinary folk perfectly well as the referents of beliefs. By saying that they are not the right sort of enduring states Ramsey et al mean they are not the right sort of enduring states by the standards of the classic model. As we shall see below, Ramsey et al make the conditional claim that *if* the most radical version of connectionism turns out to be true then the elimination of propositional attitudes will follow. This claim is only coherent, however, if FP is committed to propositional modularity. While it is almost universally agreed within the cognitive science community that this is the case (at least for those who view FP as synonymous with the classic model), it is academic arrogance in the extreme to make eliminativist claims based on this assumption. I will argue below that the cognitive science community, as well as philosophers of action, have hijacked ordinary folks' ontology and developed it into the paradigm I have been calling the PFP. The history of eliminative materialism can then be seen as the history of certain theories within this paradigm and their failings – namely the classic model. To argue for the elimination of the unembellished ontology based on the failure of its ontological elaboration is unwarranted.

The folks' folk psychology

The folks' folk psychology (FFP) may be taken to be an account of the role that folk believe propositional attitudes and intentional states play in the prediction, explanation, and causation of thought and behaviour, as well as the creation of reasons for action. A good place to start is by spending some time outlining our common sense conception of propositional attitudes. In what follows I will focus on beliefs since a discussion of beliefs will highlight some important points that can be generalised to the other propositional attitudes. We must be careful, at this point, to resist the temptation (to which many philosophers and psychologists succumb) to attribute to ordinary folk theories about propositional attitudes that more properly belong to the academic study of FP. The difficulty that has confronted academics in this area is that ordinary folk do not typically have much to say about the nature of beliefs. Nevertheless, we all have some deep-seated intuitions about propositional attitudes and it is these intuitions that we will consider here.

The first point to note is that propositional attitudes are embodied states. As Cohen notes: 'You answer the question whether you believe that *p* by introspecting or reporting whether you are normally disposed to feel that *p* when you consider the issue' (Cohen 1996: 266). Cohen goes on to distinguish between belief and acceptance. Acceptance according to Cohen does not have this associated 'feel'; it is rather a policy for inference. Thus, one can believe not *p* whist simultaneously accepting that *p*, but one cannot simultaneously believe both *p* and not *p*. Examples of this distinction are not hard to come by. I am, for example, happy to *accept* that quantum mechanics is true, but I do not *believe* it to be true (because of the counterintuitive consequences of the superposition principle). A second feature of beliefs is that they need not be occurrent states. If we accept, as seems right, that beliefs are characterised by a *disposition* to feel disposed toward a proposition, then we need not be experiencing, or indeed need never have experienced the associated feel, in order to be said to believe that *p*. As such one need never have considered *p* in order to be said to believe that *p*. All one requires in order to attribute a belief to another is one's own belief that, should they consider the issue, they would feel that *p*. This point extends to all propositional attitudes. When we say that Jon is embarrassed that he could not remember x's name, for example, he need not be embarrassed *now*. Rather, it is true to say that Jon is embarrassed that he could not remember x's name if Jon would feel embarrassed should he consider the issue. If Jon is no longer disposed to feel embarrassed should he consider the issue we would normally say that Jon *was* embarrassed that he forgot x's name.

It would be a mistake to attribute to ordinary folk a theory about the referents of propositional attitudes beyond noting that for occurrent states the referent is (or includes) the associated feel. So, for example, the referent of my belief that Jon is embarrassed is my feeling that Jon is embarrassed. Similarly, the referent of Jon's embarrassment is his feeling of embarrassment. As for the referent of dispositional propositional attitudes, ordinary folk don't seem to have any firm ideas and no strong intuitions. Question people about the referent of propositional memories and you are more likely to elicit some intuitive response. Any intuitive response, however, would fall far short of anything that could be described as a potentially falsifiable theory. (See Stich (1996) for arguments in support of the claim that ordinary folk do not have any firm intuitions about the referents of propositional attitude terms.)

On the folk conception, beliefs *can* support counterfactuals – had Jonathan not believed that the traffic lights had changed to red he would not have stopped the car. It is important to note that, although beliefs can support counterfactuals, it is consistent with the folk conception of beliefs that a person's actions are not determined by their beliefs. It is thus consistent with the folk conception of beliefs that a person holds a belief that *p* but does not act on it, or that their belief that *p* is superseded by other propositional attitudes. As Davidson notes, 'Beliefs and desires issue in behaviour only as modified and mediated by further beliefs and desires, attitudes and attendings, without limit' (Davidson 1980b: 217). When beliefs do issue in behaviour, it would seem that they do so (on the folks' conception) by providing grounds for reasons. These reasons are in turn cited as *explanations* for action.

It seems to me that this is about all one can say about the folk conception of propositional attitudes. Note that the account just bruited makes no reference to the *causal role* of propositional attitudes. Clearly people do attempt to explain behaviour by referring to reasons, beliefs and desires, but they remain mute on the question of whether these states literally cause their behaviour in the way that philosophers often suppose. Indeed, on the rare occasions when ordinary folk cite the cause of their behaviour, causation is attributed to the self (the 'I') and not to any antecedent events or mental states. Thus someone may exclaim, 'I just couldn't help myself, I had to...' An addict, for example, may feel compelled to alleviate their craving without thereby citing their craving as the cause of their subsequent behaviour. In non-pathological cases agency is never perceived to be or reported as being entirely subverted. Thus someone may say, 'I lit the cigarette because I had a craving that I could not resist' but never, 'the craving caused me to light the cigarette'.

I noted in the introduction (following Stich) that the theory of reference one applies to beliefs and desires will have important implications for the eliminativist conclusions one draws. If the above account of propositional attitudes is along the right lines, then it contains an implicit theory of reference that effectively grants immunity from elimination to the folk conception of propositional attitudes. Or, more accurately, the lack of a well worked out theory of reference guarantees immunity from elimination. The folk conception does not designate any physical state or entity to fulfil the role attributed to propositional attitudes. There can be no empirical refutation of the folk conception of propositional attitudes because ordinary folk typically do not hold well worked out theories that could be falsified and propose no powers, objects, or states whose existence we might question. Moreover, because each use of a propositional attitude term refers to an occurrent or potential 'feel', it does not much matter what causes or realises this feel. If it turns out that this feel supervenes on physical states that display propositional modularity then that's fine. If connectionism is true, and this feel supervenes on some activation pattern that does not display propositional modularity, that's fine too. Since ordinary folk do not claim to know what the subvenient bases of propositional attitudes are, that is to say what their physical referents might be, they will happily allow science to fill in the gaps in their ontology. Our stock of propositional attitude terms may expand or contract if we learn to discriminate subtle differences in our experience of holding propositional attitudes, but it is inconceivable that we have misidentified the experience of, for example, lusting after *p*, being embarrassed that *x*, etc. Even if our worst case scenario were realised and no propositional attitudes turn out to be causally relevant, this would still not entail elimination. As long as people are disposed to feel that *p*, be embarrassed that *x*, etc., propositional attitude terms will have a referent and they will not be eliminated from the popular lexicon. Similarly, though people often have some firm intuitions regarding the nature of the self (particularly if they belong to a religious group), such accounts rarely designate an entity or substance that science might find absent. Even people who believe in an immaterial soul rarely hold any theories about its interaction with the body (such as Descartes' famous theory concerning the role of the pineal gland) that might be falsified by science. Moreover, the self to which agency is attributed is known through experience and, though I argue that the experience is epiphenomenal (which robs the experiential self of its agency), the existence of the self defined in terms of the content of one's experience is indubitable. The future of propositional attitude terms within academic disciplines that seek to explain and predict behaviour by citing its cause is less secure and it is to this that we shall now turn.

The philosophers' folk psychology

What I have termed the philosophers' folk psychology is really a paradigm rather than any concrete set of theories. By this term I mean to include all those ontological elaborations of the FFP that have emanated from the philosophy of action and the cognitive sciences. The PFP goes well beyond anything that is manifest in experience and as such, unlike the common sense psychology used by ordinary folk, the theory is open to refutation by both empirical evidence and philosophical argument.

We are repeatedly told that FP allows us to predict and explain others' actions by both eliminativists and the defenders of FP. The philosophical literature is full of banal and rather contrived examples of how knowledge about a person's mental states enables us to predict their behaviour. The following account is typical of those found in the literature:

> ...if someone desires that p, and this desire is not overridden by other desires, and he believes that an action of kind K will bring it about that p, and he believes that such an action is within his power, and he does not believe that some other kind of action is within his power and is a preferable way to bring it about that p, then *ceteris paribus*, the desire and the beliefs will cause him to perform an action of kind K. (Horgan and Woodward 1991: 149)

The ubiquitous example is the action of putting up an umbrella and it is worth rewriting the above account for this specific case:

> Jon looks out of the window and notices that it is raining. This causes Jon to come to believe that it is raining. If Jon desires to stay dry, and this desire is not overridden by other desires, and he believes that taking an umbrella with him will bring it about that he stays dry, and he does not believe that some other kind of action is within his power and is a preferable way to bring it about that he stays dry, then *ceteris paribus*, the desire to stay dry and his belief that taking an umbrella with him will bring it about that he stays dry, will cause him to take an umbrella with him.

It is important to note that, in cases such as these, the occurrence of beliefs and desires are *hypothesised* to account for what appears to be purposeful behaviour. When I perform such trivial actions as getting a beer from the fridge, putting up an umbrella, or turning on the light, I am not aware of the kinds of belief-desire combinations that Horgan and others would have us believe cause my actions. Indeed were I aware of these kinds of belief-desire combinations for all my actions I am sure that I would suffer cognitive overload – my mind would become so paralysed by all these belief-desire combinations that action would become impossible. These examples, therefore, go well beyond what can be inferred from experience and what looks at first sight like an innocent description of human cognitive processes is in fact heavily theory laden. Philosophical and psychological theory is evident in two respects here, it assumes the existence of beliefs and desires that are not manifest in experience (they operate unconsciously according to the theory) and it gives those beliefs and desires a causal role. It is this putative causal role that differentiates the PFP from the FFP. As soon as one accords propositional attitudes a causal role, propositional modularity naturally follows. For Jon's belief-desire combination to literally cause his behaviour, for example, it must be the case that his non-relevant beliefs, that for example water boils at 100°C and the lawn needs mowing, are causally inactive. The two elements that define the paradigm of PFP and differentiate the PFP from the FFP then are (i.) on the PFP but not on the FFP propositional attitudes are assumed to literally cause behaviour and (ii.) in order to fulfil this role they must be functionally discrete states.

Eliminativists and defenders of FP, of course, differ on their interpretations of why talk of beliefs and desires is useful. Eliminativists (and instrumentalists) typically argue that FP is merely a generalised description that will suffice for day-to-day life but has no ontological depth. I.e. beliefs and desires do not designate *sui generis* real, causally potent properties, states, entities or anything else. Defenders of FP, in contrast, typically explain its success by arguing that at least some of the posits of FP refer to causally efficacious states.

We will examine these arguments a little later in the chapter, at this point I want to question the assumption that FP is a useful tool for prediction and explanation. The first point to note is that ordinary folk do not typically explain their behaviour in ways that parallel the causal explanations of behaviour provided by the defenders of FP (such as those in the above examples). This should not be taken to mean that ordinary folk are not adept at predicting and explaining behaviour, for clearly they are, I merely want to note that they do not typically cite folk psychological categories when they do so. Nor do ordinary folk claim to be applying a theory, invoking causal laws, or anything much else when they predict others' behaviour. Thus, if the PFP is correct, it must describe processes that are (generally) unconscious. When propositional attitudes are used to explain behaviour it is typically in cases where the behaviour was in some sense strange, atypical, or when more information is required for the behaviour to seem to 'make sense'. We should allow ourselves to be struck by how little we have recourse to folk psychological explanations and how, most of the time, our own and others' behaviour seems completely natural – without requiring explanation or justification. Doing so helps to counter the intuitive appeal of the PFP model and allows us to question why we assume that the accounts that allow us to make sense of others' behaviour parallels whatever causes us to view the behaviour of others as natural or without need of explanation. Dennett explains this as follows:

> We are *communal* folk psychologists, who are constantly explaining to other people why we think that so and so is going to do such and such. We have to talk; and when we talk, because life is short, we have to give an edited version of what we are actually thinking; thus what comes out is just a few sentences. Then, of course, it is only too easy to suppose that those sentences are not mere edited abstractions or distillations from, but are rather something like copies of or translations of the very states in the minds of the beings we are talking about. (Dennett 1991a: 142)

There are, of course, many PFP theories that fall under the broad rubric of FP. The central theme of most accounts is that FP is a theory. Moreover, it is a theory that explains how it is possible for people to make predictions because it is the theory that people apply when they make those predictions. In order to deserve the name of FP it seems plausible to suggest that the theory must posit the existence of propositional attitudes and attribute to them a role in causing behaviour. Propositional attitudes, in other words, must be the vehicle of thought and not its consequence. To put this another way, it must be the case that my experience of the desire for a beer, and my belief that there is a beer in the refrigerator (or the neural states upon which these experiences supervene), literally cause my going to the refrigerator to get the beer. Note that this is a stronger claim than that propositional attitude terms serve to explain my behaviour to others by making my behaviour intelligible to them (which does not entail that the explanation cites the cause of my behaviour). This weaker claim still gives propositional attitudes a real job to do, but here their job is explanatory not causal. An explanatory role may not satisfy theorists of mental causation but it is worth remembering that there are many perfectly adequate explanations in FP (the FFP that is) that are not causal explanations. Consider, 'Why is he in a bad mood?' 'He just got out of bed on the wrong side this morning.'

As we have already seen, propositional attitudes need not be occurrent states and a crucial aspect of PFP is that propositional attitudes literally cause behaviour regardless of whether or not they are experienced. To clarify what is entailed by this point, suppose that John goes to the refrigerator to get a beer. Prior to embarking on this course of action John thought to himself, 'I am going to get a beer from the fridge.' On the PFP model we can attribute to John a desire for beer and a belief that there is a beer in the refrigerator. This belief-desire combination was (*ceteris paribus*) casually sufficient for John's action. Now suppose that Jonathan performs the same action (the action of going to his fridge to get a beer) but without any conscious intentional beer-thoughts. According to the PFP model Jonathan's action must also have been caused by a desire for a beer and a belief that there is a beer in the refrigerator (albeit a non-conscious belief and desire) and Jonathan's belief-desire combination was again (*ceteris paribus*) casually sufficient for his action. Moreover, to explain or predict John and Jonathan's actions the PFP model would have us apply a law to the effect that if a person desires a beer, believes that there is a beer in the refrigerator, and they hold no other beliefs or desires that supersede their beer-desire, then *ceteris paribus* they will go to the refrigerator to get a beer.

In what follows we will consider the various ways that such a story could be fleshed out. Specifically, we will consider how one is able to predict John and

Jonathan's actions. Do we, for example, apply an explicit law of the type cited in the example, was the prediction made by the application of general rules or by some combination of the two?[4] We will also consider whether this account presupposes that propositional attitudes are semantically interpretable, functionally discrete states that are causally potent on some occasions and causally inert on others. This triad, first proposed as entailed by the folk psychological conception of propositional attitudes by Ramsey, Stich and Garon, has been the focus of much debate. Ramsey et al argue that, if connectionism turns out to be true, it will falsify this conception of propositional attitudes and elimination follows.

As Ramsey et al note, there are a variety of different connectionist models on offer. Some of which have units, or aggregates of units that one could plausibly argue represent propositions and function as causally discrete units. However, in what follows I will assume that something like the most radical version of connectionist models turns out to be true (i.e. turns out to be an accurate model of human cognitive processing at both the psychological and implementation levels). Since I intend to argue (with a few caveats) that the PFP (*contra* Ramsey et al) is consistent with even the most radical connectionist models, my case can only be strengthened if it turns out that neural networks do not display all the properties attributed to connectionist networks.

[4] Another possible explanation is off-line simulation. According to off-line simulation theory we predict other's actions by feeding the propositional attitudes that we attribute to them into our own 'off-line' decision making apparatus and derive a prediction as an output. Since the empirical evidence seems to be against off-line simulation, and because there are some strong arguments in support of the claim that an ability to perform off-line simulations would itself be dependent on having internalised a theory, I will not discuss off-line simulation here. For a critique of off-line simulation theory see Stich and Nichols (1996).

Stich and Ravenscroft (1996) have provided a useful summary of the forms that FP could take and their potential for elimination. Their account begins by differentiating between internal (based on Lewis' (1970; 1972) accounts) and external accounts of FP. On the external account FP constitutes a theory that systematises the 'platitudes' and generalisations to which everybody (or almost everybody) in a culture accepts. On this, external account, FP is not the type of thing that could be true or false and hence is not a candidate for elimination (what I have termed the FFP falls into this category). Since we are concerned with the PFP model here, and we have accepted that the PFP is a theory that claims to refer to the causes of behaviour (which are necessarily internal) this external account is not of much interest to us here. There is, however, an external account of FP, proposed by Greenwood, that is worth considering. Though his 'culturist' defence of FP is not quite what Stich and Ravenscroft had in mind by an external account, it is plausibly construed as such.

Greenwood's account is based on the now familiar distinction between movement and meaningful action. Greenwood argues that social psychologists (but we may extend his argument to cover the social sciences generally) are typically not interested in the *cause* of behaviour, but in how action is interpreted as meaningful social behaviour. Thus, according to Greenwood, even if it turned out that FP was wrong most of the time about the cause of our actions, or even all of the time, this would still not warrant the eliminativist conclusions that Churchland and others have drawn.

> The classification of an action as an instance of aggression or dishonesty does not presuppose any of a possible variety of competing causal explanations of aggression or dishonesty. There is, for example, no inconsistency in supposing that some instances of aggression are best explained in terms of motives of revenge, that other instances are best explained in terms of exposure to "violent stimuli"…, and that others still are best explained in terms of excitations of the lateral hypothalamus, because the correct description of an action as an instance of

> aggression does not presuppose any of these competing causal explanations of aggression. (Greenwood 1991b: 73)

The common eliminativist response to this defence of FP is to note that much of FP's characterisation of action is dependent on the intentions the agent had in executing those actions. Thus, an action is altruistic if the agent's intentions were to perform an action that would be beneficial to others but at a cost to himself. Similarly, an action is aggressive if it is caused by the intention to cause harm. Greenwood's external account of FP may save its ontology, but it is at the expense of the very purpose to which that ontology is typically applied. Greenwood's tactic here is to move the emphasis from intentions to the way that people represent their actions:

> In attributing shame to another, I mean that that person represents some actions of his or hers (intentional object) *as* personally humiliating and degrading (intensional content). Such attributions are true if and only if agents represent particular aspects of reality in the particular ways attributed to them, *irrespective of the adequacy of causal explanations referencing such psychological states.* (Greenwood 1991b: 77)

This reply presupposes the accuracy of one's self knowledge in the identification of phenomenal states, something that has been questioned in recent years. Churchland's reply to this type of tactic is to assert (in a Sellarsian fashion) that FP is committed to the view that the identification of phenomenal states is itself theory governed. And that if the theory that informs the identification of these states is mistaken, then so too is the ontology upon which it relies (see Churchland 1998). Once again Greenwood's externalism comes to the rescue here. Greenwood argues that the identification of phenomenal states does not rely on the application of a theory, but on learning how to represent *external* social reality. In the case of shame, for example, Greenwood argues that identifying shame is not a matter of discriminating internal states. Rather, it is a matter of learning to treat certain actions as 'personally degrading and humiliating'. Thus, someone can be said to have learned the concept of shame not when they can identify an internal state but when they can 'articulate this from of representation of social reality' (ibid. 83). They need not know, according to Greenwood, that it is conventionally classified as shame in their form of social life. The circularity in this argument should be glaring. The point of Greenwood's argument was to account for the self-knowledge of shame in a manner that was not informed by a theory. Greenwood attempts to redefine shame, not in terms of the self-identification of an internal state, but rather by the way that one learns to represent certain actions as degrading and humiliating. The question we must now ask is how exactly is one to identify actions that are *degrading* and *humiliating*? Degradation and humiliation are internal states which eliminativists claim, if FP were true, would require the application of a theory to identify. It is incumbent on Greenwood to show how we can learn to represent external reality without having to self-identify *any* internal states, and this is something that Greenwood has singularly failed to do.

Nevertheless, Greenwood is certainly right to redress the balance by asserting that at least some folk psychological categories are identified by reference to external reality (an aggressive act, a helpful act, an honest act, etc.). Whether or not this will grant these categories immunity from elimination, however, remains to be seen. In particular those who adopt this line of defence owe us an account of how these categories become socially meaningful. In the case of shame, for example, I argued that although Greenwood is correct in his assertion that self-identification of shame is dependent on learning to represent external reality, one can still not self-identify shame without reference to internal states. There may be examples of folk psychological categories that are not dependent on the self-identification of internal states, but they are typically not the focus of the eliminativists' critique. Moreover, even if it turns out that there are folk psychological categories that can be identified without the self-identification of internal states (identifying another person's behaviour as helpful or honest are possible examples), these categories may still prove to be candidates for elimination. The trouble is that these states may not survive the elimination of the internal states that are identified as the cause of behaviour by FP. Thus, for example, if it turns out that no actions are literally caused by revenge, hatred, or a desire to cause harm to another, we may very well choose to abandon the classification of actions as aggressive – perhaps replacing it with a more neutral term such as harmful. This is one of the categories of action that, if epiphenomenalism is true, may see the elimination of folk psychological terms from causal explanations. One might see, for example, the growth of pathological explanations for criminal behaviour as a precursor to the eventual elimination of folk psychology from jurisprudence.

Stich and Ravenscroft identify a number of elements that an internal account of FP might include. The first way that folk psychological capacities might be subserved is by a store of information (possibly in the form of propositions or declarative sentences) that can be applied to guide behaviour. A second way to construe FP is some combination of both rules and information. In this case, for example, the rules might function as a guide for the application of the information (but note that there are many possible ways that rules and information might interact). The final possibility is as a Fodorian style set of internalised rules, which might be either learned or innate, consciously accessible or unconscious/tacit, that are applied as guides for problem solving, to accomplish tasks, predict others' behaviour and so on. On Fodor's account there need not be any 'central meaner' to understand the rules of the tacit 'theory'. Rather, each rule can be broken down into simpler rules which are, in turn, composites of 'elementary operations' that can be performed unthinkingly by the nervous system (see Fodor 1968).

Stich and Ravenscroft argue that only the former two options are potential candidates for elimination. If the latter option turns out to be the correct version, that the system that subserves our folk psychological capacities (of prediction and explanation) consists of all rules and no propositions, then it will make no sense to claim that FP is either true or false. Thus Stich and Ravenscroft argue that eliminativists who adopt an internal view of FP must be committed to the view that FP isn't all rules and no propositions (129-30). This conclusion would only be warranted if the rules made no reference to propositional attitudes. However if this were the case then the eliminativist would win by default since any system that does not make use of propositional attitudes hardly deserves the name of *folk* psychology. On the other hand, assuming the rules relate in some way to propositional attitudes, eliminativism is still a viable option. All the eliminativist need do is show that propositional attitudes do not exist and they will get the elimination of the rules that relate to them for free. If this were the case, and the eliminativists get their way, then paradoxically it would guarantee causal relevance for propositional attitude terms. This surprising conclusion is entailed by the fact that we know humans have the ability to predict one another's behaviour. Thus, if these capacities are subserved by a set of rules that make reference to propositional attitudes terms, it would be in virtue of our use of propositional attitude terms that we are able to predict behaviour – *ex hypothesi* propositional attitude terms would be causally relevant.

Conclusion

I want to conclude this chapter by returning to the concept of causal relevance and by exploring how our experience of believing, desiring, etc. fits in with supervenience and connectionism. Before doing so it is worth taking stock of the ground covered so far. Starting with the classic model, there seems little doubt that if the brain turns out to be a connectionist network then cognitive thought cannot consist of rule governed symbol manipulation and the classic model must be false. I have not spent much time defending this claim because it seems to be generally accepted and because the truth or falsity of the classic model has little bearing on the social sciences whose methodology is based on something more akin to the PFP. The FFP survives untouched by connectionism, I argued, because it does not make the sort of empirical claims that could be falsified by connectionism or any other empirical theory. Where the FFP does make 'empirical' claims they are in the form of generalisations and near universally accepted platitudes (what Stich and Ravenscroft characterise as an external model) and hence cannot be falsified by connectionism. This leaves us with the PFP. The PFP, since it purports to be about the causes of people's behaviour, is certainly an internal account by Stich and Ravenscroft's standards. As such it is the type of theory that could be falsified. In what follows I will argue that the PFP can survive the connectionist onslaught but not entirely intact. Specifically I will suggest that the PFP is only an adequate causal-explanatory model for those cases where behaviour is caused by occurrent states.

Let's start by stating some uncontroversial observations:

1. Cognitive thought is experienced (at least some of the time) as verbalised (i.e. as a series of words, expressions, and sentences). Propositional attitude terms

frequently occur in this verbalised thought. However we often also experience propositional attitudes without verbalising the experience – we can hope, desire, believe, want, and so on, without verbalising these experiences.

2. When we communicate with others we do so using language and propositional attitude terms form part of that language.
3. (From 2) If our brain is a connectionist system then part of its input and output is in the form of language, propositional attitudes, declarative sentences and the rest.

To this list we may now add a fairly uncontroversial premise (at least for materialists):

4. Our internal dialogue containing propositional attitude terms, and the experience of propositional attitudes themselves, supervenes on synchronous internal states.

If premise 4 is correct then it would seem to rule out the wholesale elimination of propositional attitudes and lend some support to the PFP since:

5. *Ex hypothesi,* there must be subvenient neural states that physically realise the experience of holding propositional attitudes and using propositional attitude terms. Consequently, there must be some neural state one could identify as the physical realisation of the use of any given propositional attitude term and the experience of propositional attitudes themselves (even if those states are merely activation patterns rather than enduring states of the system).

Connectionist motivated eliminativists would not necessarily deny this conclusion. What eliminativists would deny is that cognitive processes operate by the manipulation of these aforementioned states (in the manner of the classic model). Nor would connectionists deny point 3, that if the brain is a connectionist system then its input and output is in the form of language containing propositional attitude terms. Rather, connectionist motivated eliminativists would claim that the interesting part of cognitive thought, the bit between the input and the output, does not contain anything that could be identified with propositional attitudes, let alone anything like the rule governed manipulation of those propositional attitudes. This is the crux of the matter for the PFP, for if there is nothing in the brain that could be identified with beliefs then it cannot be true that our beliefs cause our behaviour.

The eliminativists' claim, however, is only plausible by the strict criteria of the classic model. If the brain is a connectionist network then it doesn't look like we will find any *enduring* states that display propositional modularity and the classic model will have been proved wrong. The PFP model, however, does not typically make any specific claims about whether or not dispositional beliefs are enduring states that display propositional modularity. Thus, where dispositional beliefs are invoked in explanations of behaviour, a system's disposition to realise an activation pattern that could be identified with an occurrent realisation of the dispositional belief in question would be compatible with most versions of the PFP. Beliefs in this case, however, could not be said to literally cause behaviour. Rather, neurally realised dispositional beliefs form part of the necessary background conditions for action. Where beliefs are invoked in causal explanations all that is required to vindicate the PFP is that beliefs display propositional modularity at the point when they cause action and causally influence cognitive thought, and activation patterns will adequately fill this role.

When behaviour is explained by non-conscious beliefs, however, we have no basis for inferring the existence of activation patterns that fulfil these criteria. As I noted during my discussion of the PFP, the existence and causal efficacy of non-conscious propositional attitudes are hypothesised by PFP to account for seemingly purposeful behaviour. For occurrent states supervenience entails the existence of a physical base and allows us to make certain inferences about the properties belonging to this physical base. No such inferences can be made for dispositional states or states whose existence are merely hypothesised. This has the unfortunate consequence that if it can't be explained by folk then it can't be explained by the PFP. Since academia hopes to go beyond the common sense explanations of ordinary folk, it will have to look beyond the PFP to find its explanations.

Returning now to causal relevance, in chapter 2 we accepted Kim's multiple realisation thesis and the accompanying hypothesis that mental states supervene on physical bases with disjunctive causal powers. The substance of this hypothesis was that the causal power of mental content is identical to that of its subvenient base, but that the same mental content may be realised by a variety of different bases with diverse causal powers. Thus, if the physical realisations of beliefs are activation patterns, the causal powers of those activation patterns are likely to be disjunctive – on some occasions they may be sufficient (ceteris paribus) for action and on others neither necessary nor sufficient. Our final conclusion then is rather sketchy. Neither our best case nor our worst case scenarios of mental relevance for occurrent states are quite right and the truth must lie at some indeterminate point between the two. Or, perhaps, at different points for different actions, people, times, and so on. For dispositional propositional attitudes connectionism seems to entail something like our worst case scenario. Although it is consistent with connectionism that dispositional propositional attitudes (what the PFP would claim are non-conscious beliefs and desires fit into this category) support counterfactuals, we are not entitled to claim that they literally cause behaviour. If the brain is a connectionist network then to refer to non-conscious/dispositional propositional attitudes is to adopt the intentional stance. Although the intentional stance may be a useful predictive tool it has to be recognised that it cannot supply causal explanations.

Chapter 5

Epiphenomenalism

In chapters 2 and 3 we reviewed all the metaphysically possible solutions to the problem of mental causation (or at least all those materialistically kosher positions) and found none that are compatible with the three principles outlined in the introduction – except for epiphenomenalism that is. Faced with a choice between giving up one or more of our three principles or our long standing commitment to mental causation, I have opted for the latter. Epiphenomenalism, however, is a most unsatisfactory solution which, as we shall see, saddles us with just as many problems as it solves. The thought that I, along with the rest of humanity, am nothing more than a conscious automaton is not one that I enjoy and in truth it is not one that I believe (in the sense of belief versus acceptance outlined in chapter 4). The consequences of epiphenomenalism are so counter-intuitive that I think we would be justified in questioning the sanity of anyone who claims to believe it. As Kant pointed out no determinist can *choose* to sit back and wait to see what their beliefs and desires will cause. Similarly, no epiphenomenalist can choose to ignore their causally inefficacious phenomenal states and wait to see what actions their brains will instigate next. The human constitution, it seems, is not capable of embracing either determinism or epiphenomenalism as a lifestyle choice. Nevertheless, when I face the intellectual choice between giving up my *acceptance* of materialism and the epiphenomenalism that I have argued it entails, or my intuitive *belief* in mental causation, materialism and its unpalatable consequences wins the day.

To recap, we (reluctantly) adopted epiphenomenalism because we found it to be the only position compatible with the following three principles:

P1 The irreducibility of phenomenal states.

P2 The causal closure of the physical.

P3 The principle of causal explanatory exclusion.

The latter two of these principles have long been accepted as the foundations of the materialist perspective and are generally regarded by those working within the materialist tradition as uncontroversial. Our first principle, which gained currency in the latter half of the twentieth century following some convincing arguments proposed by Searle, Nagel, and others, is now also generally accepted. The current incarnation of the mind-body problem (stemming largely from Davidson) concerns the apparent incongruity between these three principles and the assumption that mental states are causally efficacious. Our principal goal in part I was to review the various attempts to reconcile these principles with the assumption of mental causation. I have argued that, to date, all these attempts have violated one or other of the three principles. Moreover, I suspect that we have now exhausted our stock of metaphysically tenable positions. It will be objected that even if one accepts the conclusions of part I (that all previous attempts have failed) the future state of knowledge cannot be predicted and a satisfactory solution may still be forthcoming. We have seen, however, that the devil has not been in the detail. Rather, the detail has served to obscure some fundamental and irreconcilable inconsistencies in the materialist position. We can, to be sure, expect many more sophisticated theories that putatively demonstrate the causal efficacy of mental states. Ultimately, however, any new theory will face the ineluctable choice between non-reductive and reductive physicalism, assert mental-physical covariance or divergence, and propose 'upward' or 'downward' causation. Our survey of supervenience and emergence has taken us back and forth within these parameters and we have found no position consistent with the three principles set out in the introduction. That is not to say that progress had not been made in recent years. The supervenience thesis in particular has arguably secured the causal efficacy of mental states that can be functionalised (such as mental content) and can, or so I argued, be used to demonstrate the causal relevance of phenomenal states. However, as I have stressed repeatedly, phenomenal states resist functionalisation in the manner of mental states that can be individuated independently of their PE (as in the case of the content of propositional attitudes).

Having exhausted our stock of metaphysically tenable positions we face a stark choice, physicalism and PE causation seem irreconcilable and something will have to give. Ultimately we must choose between abandoning our commitment to one or more of our three principles, or retaining our commitment to these principles and accepting epiphenomenalism. This is uncharted territory for philosophy and we have few principles to guide our choice. Both positions are deeply offensive to our intuitions and ultimately the choice is likely to depend on which theory's consequences are more in accord with our worldview. If we choose the former option, we may retain our belief in mental causation but at the expense of our scientific worldview (since this option will probably entail some form of dualist interactionism or the radical sense of emergence involving emergent powers). If we choose the latter then we solve the problem of mental causation, or more accurately we avoid the problem. Nevertheless, this option generates some philosophical and epistemological problems of its own and it is to these that we will now turn.

In this chapter I want to explore the concept of epiphenomenalism in a little more depth. Though nothing I will say in this chapter affects our previous conclusions this exercise will help to elucidate some of the problems that epiphenomenalism generates and to contextualise these problems in relation to those of non-reductive physicalism. I also hope to question the basis of our dismissive attitude towards epiphenomenalism by contrasting our counterintuitive reactions to its consequences with our counterintuitive reactions to the consequences of non-reductive physicalism. My aim is to show that epiphenomenalism is no better or worse off than contemporary non-reductive physicalism and deserves to be given serious attention.

Evolutionary arguments against epiphenomenalism

Evolutionary theory presents epiphenomenalism with its greatest embarrassment. Having left qualia dangling, we have denied ourselves access to the standard Darwinian explanation for their existence. The critique of epiphenomenalism, which barely receives a paragraph in most texts, is always of the same form. First there is the dogmatic assertion that we all know that mental states are causally efficacious. This is followed by the more sophisticated argument that qualia aren't brute facts of nature.[1] Rather, they are the consequence of a long evolutionary history and that they evolved precisely because they are causally efficacious – endowing their possessor with an adaptive advantage. It is then claimed that one of the driving forces of evolution is efficiency, and that mutations that are high in energy consumption tend to be weeded out unless they confer some evolutionary advantage on their possessor. It is further noted that the

[1] The only exception to this argument comes from panpsychists. Chalmers being the best known contemporary advocate.

brain consumes a proportionally vast amount of energy and that much of the activity of the brain is directly involved in consciousness. The conclusion is that the very existence of conscious mental states serves as compelling evidence for their causal efficacy. Although there are cases of so called 'spandrels' in nature, they are typically low in energy consumption. If qualia are mere spandrels then their intensity and degree of complexity is astounding, prima facie they are too big for the building to support![2]

Those who adopt this argument are quite right to note that nature tends towards frugality. The trouble with their argument, however, is that it does not differentiate between the energy required to generate conscious experience and the energy required to generate the neural correlates of conscious experience (identity theorists would, of course, get around this problem by claiming that the experience and its neural correlate are identical). We do know that much of the activity of the brain is related to the performance of tasks with which consciousness is associated. However, until we know what (if any) causal contribution is made by the experience we will be unable to decide if conscious experience is too expensive to hitch a ride on the brain's evolution. If PE were some property of the functional organisation of the brain or of certain biological processes, rather than being some distinct 'substance', then no extra energy would be required in order to bring it into existence. Indeed I will argue that a persuasive case can be made that not only is consciousness not too biologically expensive to hitch a ride on the brain's evolution but that it comes at no extra

[2] Spandrel is an architectural term which denotes an ornamental work (that performs no structural function) designed to fill the gap between the curves of arches and its surrounding rectangular mouldings. Gould and Lewontin (1979) borrowed the term to refer to a phenotypic trait that performs no biological function.

cost. Moreover, I will suggest that this position is entailed by most materialist accounts of mental causation.

Suppose that there is some property P of the brain in virtue of which consciousness is instantiated. P is invariably implicated in the causal chains with which consciousness is associated – it is involved in causing behaviour, it occurs as a consequence of environmental stimuli, and it is involved in causing other neurological states that also have property P. Suppose also that P consumes the lion's share of the brain's energy supply. Even if we knew all there is to know about P, I suggest, there is no evidence on the basis of which we could infer that consciousness is biologically expensive. I will illustrate this with an example. Suppose that all members of the genus *car* are conscious and that there is some property P of the engines of cars in virtue of which cars are conscious. Some species, it turns out, have much more complex conscious lives than others. The Smart car is only minimally aware of its environment while the Jaguar XKR has a sophisticated understanding and awareness of traction control, aerodynamics and sound proofing (amongst other things). Mechanics, we might imagine, will one day discover the property that is instantiated in the engines of cars in virtue of which they are conscious. Suppose it turns out that it is the property of being a piston chamber that instantiates consciousness in cars (is the property in virtue of which cars are conscious). Smart cars with their meagre 698cc engines consequently have much simpler conscious lives than Jaguar XKRs with their massive 4.2 litre V8s. Piston chambers consume almost all of the energy available to cars since it is there that combustion occurs. No mechanic in their right mind, however, would claim that the property of being a piston chamber is too expensive to hitch a ride on the evolution of the car. Though there are many other mechanisms that could have performed the piston chamber's function more efficiently (perhaps electric motors might have been more efficient), being a piston chamber is an absolutely indispensable evolutionary trait for the species Jaguar XKR. Notice that all the energy consumed by the piston chamber is accounted for – some is lost as heat, some turned into kinetic energy, and so on. There is no energy left over to produce car-consciousness. *Ipso facto* if you belong to the genus car and you have a piston chamber you get consciousness for free.

An exactly parallel argument can be constructed for the property P of the brain that instantiates consciousness, *whatever that property turns out to be.* No matter what property or properties instantiate consciousness in humans – whether it is functional, relates to biological hardware, occurs at the quantum or atomic level – the property must participate in a causal chain. Assuming it does so then, just as with combustion in a piston chamber, all the energy consumed by the physical states that instantiate the property will be accounted for. If you are a human and have P you get consciousness for free.

If one accepts, as I do, that PE is irreducible then this is, I think, a compelling argument against contemporary materialist theories of mind. To pursue the above analogy, any microphysical states or effects of the piston chamber must be caused, according to materialism, by antecedent microphysical states. As such there is no causal work left over for car-consciousness to perform. If conscious states have causal powers they must utilise some energy in exercising those powers (again according to materialism). Assuming that all the engine's energy supply is accounted for by microphysical causal relations, however, there is no energy left over for consciousness and car-consciousness must be epiphenomenal. Those who view the evolutionary argument against epiphenomenalism as compelling had, for that reason, better look towards emergence for their theory of mental causation. Only if consciousness exhibits emergent powers would the presence of consciousness result in greater energy consumption than that of its physical realisation alone. Materialism will not tolerate missing energy or events without a cause since both would be a violation of P2. If such events were to occur then materialism would be false and some version of emergence or dualism would be vindicated.

The evolutionary argument against epiphenomenalism, however, is a double edged sword. We have yet to consider the argument that PE would not exist if it were not causally efficacious. I have to confess that I can neither provide nor know of any convincing counterarguments. Epiphenomenalism seems to entail that PE is just a brute fact of nature. Such a statement is not likely to be met with anything other than disdain by those who argue in favour of mental causation. It is worth remembering, however, that as a brute fact of nature PE is in distinguished company. The metaphysical question, 'Why is there something rather than nothing?' no longer prompts scientists into providing teleological explanations that cite some divine will. What's more, whenever a divine will has been invoked to explain the existence of the physical universe, it is the creator's existence that becomes the unexplained fact. Although I think that we should always continue the search for answers to such metaphysical questions, we should at least recognise the possibility that a *just so* story is all there is to tell.

Epiphenomenalism and epistemology

Although I believe epiphenomenalism is the most plausible doctrine regarding mental-physical causal relations, it contains numerous inconsistencies that make it far from ideal. Specifically, epiphenomenalism generates some serious epistemological problems relating to both self-knowledge and the problem of other minds. Epiphenomenalism, it will be remembered, states that PE has no causal influence on physical states although it is caused by them. By endorsing the supervenience thesis we have also ruled out the possibility that the PE has any casual influence on future mental states. The reason for this is that supervenience states that mental properties are dependent on *synchronous* internal physical states. Thus antecedent mental states are entirely inefficacious – they affect neither mind nor matter. Prima facie the phenomenology of self-conscious awareness seems to refute this. I am aware of phenomenal states not physical states and I could not be aware of phenomenal states unless they had some causal powers. In other words, when I react to the sights and sounds around me I react not to the neurophysiological cause of my perceptions but to the qualia themselves (or at least this is how I experience my interaction with the world). Thus, if I am waiting for a bus, I do not experience (directly) the event in the external world of the bus's arrival as causing me to think, 'Ah good the bus.' Rather, I experience my thought, 'Ah good the bus' as being caused by my having noticed a big red phenomenal object appearing in my visual field. In such cases it is the big red phenomenal object that I am aware of and not whatever is going on in my primary visual cortex (or so the OMRists' story goes).

There are two issues here and we must be careful at this point to clearly differentiate them. The first relates to the putative causal powers of the big red phenomenal object. The reason we were persuaded by epiphenomenalism is that, as physicalists, we believe that there must be a complete causal account that explains our noticing the arrival of the bus which does not cite phenomenal states. Such a story might include such things as the surface of the bus reflecting light waves with a frequency of 600 nanometers, light hitting the retina, nerve impulses passing down the optic nerve and the neurophysiological events going on in the primary visual cortex, but would make no mention of the big red phenomenal object (or quale). Since we have rejected the idea that the big red quale might be identical to any of the things mentioned in our story, it follows that the big red quale is epiphenomenal. Thus, although it appears to us that we are reacting to qualia, if epiphenomenalism is true then qualia do not from part of the causal chain that links perceptions and reactions. I use the term perceptions here (and in what follows) to refer to our ability to acquire true beliefs about the world. In the context of the foregoing discussion I take it that this ability is subserved by neurophysiological events, with the conscious experience of qualia being their epiphenomenal consequence. Thus, by stating that we react to our perceptions, I mean that our behaviour is caused (in part), not by our reacting to qualia, but by the antecedent neurophysiological events that subserve perception.

The real problem here relates not to the putative causal powers of qualia on action and thought but to the epistemological question of how we can come to be aware of our phenomenal states at all. In the above story, when the bus rounds the corner, I not only experience a big red quale but I know that I am experiencing a big red quale. The trouble with this is that if phenomenal states really are epiphenomenal then we have to ask ourselves how we know that they exist. In other words, how can we know of the existence of a property that has no causal powers? Clearly though, we are all aware that phenomenal states exist, we can predict their occurrence, philosophise about their role in mental causation and alter the physical world by (*inter alia*) writing books about them. Epiphenomenalism, therefore, is prima facie self-refuting. As an epiphenomenalist, of course, I am committed to arguing that our self-conscious awareness of our conscious states is also an epiphenomenal consequence of physical events. It seems counterintuitive, however, to claim that the subvenient base for self-consciousness would exist in the absence of the PE of self-consciousness. It is not hard to imagine the evolution of complex neural structures, such as those that subserve pain behaviour (we will call these structures physical pain), in the absence of the phenomenal experience of pain (phenomenal pain for short). That is to say, I can happily accept that it is a contingent fact that physical pain causes phenomenal pain and that the evolution of physical pain would have occurred in just the same way regardless of whether physical pain had caused or realised phenomenal pain. To return to the distinction between belief and acceptance, as an epiphenomenalist I am happy to *accept* that the evolution of the brain was unaffected by the presence of PE. In the case of physical and phenomenal pain I *believe* that physical pain would have evolved in just the same way regardless of the presence or absence of phenomenal pain. However, when I consider *all* PE, and self-consciousness in particular, I no longer find it intuitively plausible (that is, I no longer *believe* it to be possible) that the brain's evolution would have occurred in just the same way without PE. Moreover, it seems even more incredible that the ontogenetic development of the brain is unaffected by the presence of PE. Although I accept that epiphenomenalism is true I cannot *believe* that a molecule for molecule identical world populated by automatons could contain philosophers who worry about the causal efficacy of non-existent qualia. To put this in the starkest possible terms, as an epiphenomenalist I am committed to claiming that had I been born an automaton with no conscious experience I would still have written a thesis entitled *The Epiphenomenal Mind*.

It is easy to be seduced by these sorts of worries into conceiving of the brain as some form of Cartesian theatre upon which our mental lives are played out. It seems that the brain must somehow be 'aware' of the existence of PE in order to account for the possibility of self-knowledge and the fact that discourse about PE forms part of the physical world. The idea of a Cartesian theatre, though, only serves to complicate matters without solving any of the fundamental problems. Nevertheless, it is worth thinking through what such a Cartesian theatre might look like since this will help to elucidate the problems faced by epiphenomenalism. To that end I present the following analogy which also helps to clarify some of the problems faced by non-reductive physicalism.

Suppose that some scientist designed a conscious robot, we will call it Art after Davidson's robot in *The Material Mind*. Now suppose that Art is fully self-conscious and experiences the world much as we do. He is in possession of fully-fledged propositional attitudes, experiences pain and pleasure and even has a theory of mental causation. However, Art's mental states are epiphenomenal, they are caused by events in Art's circuitry but they have no causal influence on either Art's circuitry or on his future mental states. Suppose also that the scientist, we will continue the theme and call him Donald, is in complete control of Art's actions – perhaps via some handset such as those used to control computer games. Now suppose that Donald has had a long standing interest in cybernetics and he has figured out a way of transmitting and translating the electrical patterns upon which Art's conscious states supervene directly into his own nervous system. Much to Donald's delight he found that he was able to experience the world exactly as Art does, he experiences the pain that Art feels when he skids into a wall and the pleasure that Art feels when he is oiled and his batteries recharged. In this scenario we will further imagine that Donald is still self-conscious and retains complete control of Art (via his computer console). If this fantasy scenario were to transpire then we can assume that Donald would do everything in his power to ensure that Art experienced only pleasurable sensations. Donald would take great care, when playing with Art, not to lose control and let him skid into walls and he would ensure that Art was well maintained, oiled, and his batteries recharged regularly.

If we allow the analogy to be stretched further into the realms of science fiction, we might imagine that Donald can become so absorbed with his new toy (in the manner that a child might become engrossed in a computer game) that he temporarily loses all sense of self. He begins to experience the world *as* Art. Just as a child might become so engrossed in a computer game that they react unthinkingly to the stimuli on the computer screen, Donald begins to react automatically to the stimuli that Art experiences. Now, instead of reacting to Art's desire to be recharged by thinking, 'I experience Art's desire for ____' Donald's unconscious brain is doing the thinking for him, when Art experiences a desire Donald's brain takes the appropriate actions to fulfil the desire.

In effect, Donald's brain has become a Cartesian theatre upon which Art's mental states are played out. Instead of the Cartesian self, however, the audience is composed of a few pounds of neurons, glial cells and the rest, arranged in such a way as to minimise Art's pain and maximise Art's pleasure. Now suppose we replace Donald, in our analogy, with evolution and Art with an 'unintended' and fortuitous consequence. We might suppose that evolution hit upon the idea of a conscious self as an indicator of the relative fitness of the organism. One of the main problems with epiphenomenalism is in accounting for the correspondence between the function and feel of mental states (i.e. why, if the PE of mental states is epiphenomenal, do harmful stimuli cause pain rather than pleasure), this analogy solves that problem by making phenomenal states an indicator of the state of the organism.

Although this analogy may strike one as (briefly) plausible we soon find that we have come full circle and face exactly the same problems that we sought to solve. To recap, we initially adopted epiphenomenalism because we could not find a place for the PE of mental states within our physicalist ontology and because we were unwilling to adopt any form of dualist interactionism. We now face exactly the same problems with the conscious robot analogy. That is to say, we are still left with the problem of explaining how the PE of Art's mental states can have any effect on Donald's brain. We got around this problem initially by making the subvenient base of Art's phenomenal states do the causal work but we must now confront the fact that we have merely added a further layer of complexity (Art's phenomenal states and the circuitry upon which they supervene) without solving any of our problems. The PE of Art's mental states is epiphenomenal. All the causal work is being done by their subvenient base. Thus in our analogy Donald is not reacting to Art's PEs but to the electrical events in his circuitry (the circuitry upon which Art's PEs supervene). Although this scenario is ridiculous in the extreme, it is worth remembering that this is precisely the set of problems that confronts non-reductive physicalism and for which they, as yet, have no solution.

Knowing one's own mind

Epiphenomenalism also forces us to confront some profound issues relating to self-knowledge, introspection, and the nature of the self. We will consider the nature of the self in part II, at this stage I want to focus on the problems epiphenomenalism generates for self-knowledge and introspection. Specifically we need to consider how one can be aware of perceptions, sensations and thoughts. As we have already touched upon above, this is a separate problem from mental causation or the problem of how the physical world can contain discourse about epiphenomenal states.

'Are you in pain?' is the sort of question that one can normally answer without the need for the type of introspection that might accompany questions such as 'Do you believe that *p*?' There are, of course, borderline cases such as the lack of concern one has for one's pain when on morphine, the ability to ignore chronic pain, or the threshold between discomfort and pain. Taking the last case as an example, if someone on the threshold of discomfort and pain is asked if they are in pain they may have to reflect on the nature of their sensation. In such cases, however, they do not discover pain that they were previously unaware of, rather, the focus of their attention serves to amplify the sensation such that it becomes painful. In such cases focussing on the sensation of discomfort may literally cause discomfort to become pain. These observations are neither original nor contentious; I mention them only to highlight the fact that PE is necessarily conscious. The second important point is that we experience a distinction between the sensation (thought or perception) and the self which is the subject of conscious experience. We do not experience the self *as* a bundle of sense perceptions and thoughts (as Hume would have it). Regardless of the way one conceives of the self, the inescapable fact remains that we draw a distinction between different thoughts, sensations, and perceptions, each of which we confront as the object(s) of our conscious awareness, and the self. Armstrong defines introspection as '...a mental event having as its (intentional) object other mental happenings that form part of the same mind' (Armstrong 1994: 109). Since, according to Armstrong, a mental event can not be aware of itself, introspection must involve what he terms a 'self-scanning' process. Thus, becoming aware of a mental state requires a second mental state to scan the object of introspection. As a means of escaping the inevitable infinite regress entailed by this argument (where we require a third mental state to introspect the second, and a forth to introspect the third...), Armstrong claims there must be an 'unscanned scanner' or self. Thus, for Armstrong, the self is conceived as that which is doing the scanning at any given time. If our version of epiphenomenalism is true then it should be obvious that this conception of the self and introspection requires modification. In our version, the job of scanning (that is becoming aware of a mental state) is not a function of the self or of any mental event. Rather, if epiphenomenalism is true, 'introspection' is a function of neural events and bears no causal relation to the PE that is the putative object of introspective awareness.

Since epiphenomenalism entails that there is no self doing the introspecting, or any mental state having as its object other mental states, we are forced to argue that phenomenal states are intrinsically 'self-conscious'. We can best approach this issue by considering the following question: Would we know that we were in pain if pain were the only thing that we experienced? We would not, of course, be able to think reflexively about the sensation – we could not say to ourselves 'I am in pain' – since this would constitute an experience distinct from the pain (a mental state having the pain as its object of introspection). Nevertheless, though we may lack the cognitive recourses to label the sensation, there are no good reasons why a state of pure pain should not be possible. If this is the case then epiphenomenalism presents us with no special problems for self-knowledge.

Clearly though, this entails that cases of putative self-knowledge are merely illusions in the sense that saying or thinking 'I am in pain' is an event completely unconnected with the PE of pain. This presents us with the following picture which contrasts with the traditional model where self-knowledge relates directly to the PE:

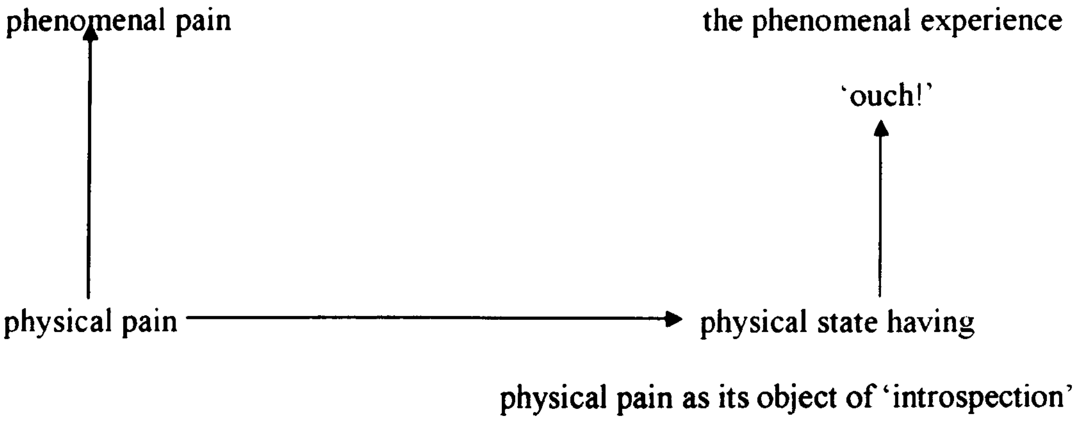

phenomenal pain
the phenomenal experience
'ouch!'
physical pain
physical state having
physical pain as its object of 'introspection'

Fig 5.1 Epiphenomenalism and self-consciousness

If this model of introspection is accurate then epiphenomenalism suggests that the problem of self-knowledge is merely a grammatical illusion generated by the use of the personal pronoun in relation to PE (as in, *I* am in *pain*). If the version of epiphenomenalism being developed here is correct then there is no *I* to experience pain. Rather, there are a succession of (directly) unconnected phenomenal experiences, phenomenal pain followed by the PE 'ouch' as in the above diagram, for example. Though few endorse epiphenomenalism, the idea that we do not have direct access to qualia, that instead we are constituted by them, is fairly common. Clark expresses this possibility as follows: 'we can't, finally, say what it's like to have qualia since we don't have a first person perspective on them: we don't 'have' them at all, neither do they 'appear' to us, nor are we 'directly acquainted' with them. We, as subjects, *exist as* them' (Clark 1998: 51). We are, of course, left with Hume's binding problem: what causes this succession of phenomenal experiences to be experienced as belonging to a single subject. However, as I will argue later in this thesis, if I am correct in advocating epiphenomenalism then Hume's binding problem is a one for the neurologists to ponder and one that I am unqualified to answer.

Ontological implications of epiphenomenalism

Thus far we have focussed on the relationship between mental and physical states in the context of causation. We have dealt with the ontological nature of the content of mental states by showing that it is reducible to (perhaps even token-token identical with) physical states in the brain – thereby guaranteeing its causal efficacy. However, we have yet to consider the ontological status of PE. The irreducibility of PE was the principal reason why we concluded that it must be epiphenomenal but its irreducibility also has important ontological implications. Specifically, we have to confront the worry that the irreducibility of PE entails that it is non-physical. To explicate, we have concluded that PE (say the PE of a pain in my foot) is caused by, but not identical with, events in the brain (in this case some event in my somatosensory cortex). The question we must now address is what has been brought into existence when an event in my somatosensory cortex causes the experience of pain in my foot. Or, alternatively, what has been modified by the event in my somatosensory cortex causing a pain in my foot. If we claim that some material substance has been brought into existence, or in some way modified, then there would seem to be no good reason why this material substance should not be capable of causally interacting with the brain region with which it is spatio-temporally coexistent. Indeed, as Popper (1977: 72) notes, the idea that there could be something which can be acted upon, but which cannot act, is likely to contradict Newton's third law of the equality of action and reaction. Though Popper applies this as a general argument against epiphenomenalism, regardless of the ontological status of the epiphenomenal property, it only holds when the epiphenomenal property is physical (after all one can hardly expect a law relating to the material universe to apply to non-material substances, if there are indeed such things).

The real problem with the possibility that PE is some material substance, however, is that it would seem to contradict our first principle. The principle of the irreducibility of PE, it will be remembered, states not that PE is irreducible to neural states but that PE is irreducible to physical states. Thus if PE were some material substance brought into existence by events in the brain it would be irreducible to those states and we would have the start of an infinite regress.

Alternatively we could embrace some form of dualism, either property or substance dualism. However, as I will argue below, I doubt that property dualism (where the two properties in question are physical and phenomenal) is consistent with epiphenomenalism. Property dualism (or double aspect theory) states that one thing, a person, can have both mental and physical properties and that neither of these properties can be reduced to the other. This is a perfectly consistent position with regard to the *description* of mental and physical properties (c.f. my discussion of Davidson's Anomalous Monism in chapter 2) but faces problems when the properties in question are intrinsic. To say that the pain in my foot is a property of a physical event in my somatosensory cortex implies that the pain is identical to the physical event (or an aspect of it), and this is something which I have strenuously denied. I experience the pain in my foot as being located in my foot (that is to say, the spatio-temporal[3] location of my phenomenal pain is different from the spatio-temporal location of my physical pain), property dualists would argue that this experience is an illusion, that the physical pain and phenomenal pain are both properties of a single physical event located in the somatosensory cortex. However, I have argued at some length (following Searle) that where PE is concerned the illusion is the reality. Accepting this has the consequence that (in this case) the pain is *literally* in my foot and cannot, therefore, be a property of an event in my somatosensory cortex.

To elaborate, one line that an epiphenomenalist might adopt is that the PE of pain is analogous to the colour properties of chlorophyll. Being green is, after all, epiphenomenal with respect to photosynthesis. Such a position has the advantage that it would not fall foul of the evolutionary argument against epiphenomenalism. Despite being epiphenomenal with respect to photosynthesis, chlorophyll's instantiating the property of being green is not a biologically expensive trait. Moreover, there is a perfectly legitimate physical explanation for the greenness of chlorophyll. The trouble is that the property of being green is (token-token) identical to the tendency of chlorophyll to reflect light waves of a

[3] The temporal aspect of this argument concerns the possibility that phenomenal pain occurs after physical pain. This is implied by the claim that phenomenal pain is caused by physical pain and

particular frequency. Greenness and the property of reflecting light waves of a specific frequency are temporally and spatially coextensive. PE, I have argued, fails this identity criterion and cannot, therefore, be identical to a property of a physical state.

This restricts our choice to an unsatisfactory version of materialism, where we postulate the existence of some material substance devoid of causal powers, or substance dualism. Although substance dualism (where the non-material substance is epiphenomenal) is consistent with our three principles – causal closure, causal explanatory exclusion, and the irreducibility of mental states – it seriously undermines the metaphysical position from which these principles are derived. If we were to accept substance dualism, therefore, we would have to at least question our commitment to the three principles, if not abandon them altogether. Moreover, it is a small step from accepting substance dualism to embracing full blown dualist interactionism. After all, if matter has the potential to generate some non-material 'substance' there is no reason to suppose that this 'substance' should not be able to causally interact with the matter from whence it sprang. To my knowledge there has never been any serious attempt to elucidate the ontological nature of non-material substances (such as Cartesian or Platonic souls, for example) in the way that physicists have done with physical substances. However, substance dualism has always been associated with dualist interactionism and this is a view which we have already rejected.

by our discussion of neuronal adequacy in ch. 2. For the purposes of this argument a spatial difference is sufficient and nothing hangs on there also being a temporal difference.

There is one more option open to us but I confess that I find it even less plausible than the first two. We could attempt to argue that PE *is* the arrangement of matter, rather than matter itself, as functionalists might want to argue. However, not only would such a move prompt the well known arguments against functionalism[4] but it would entail that PE is (token-token) identical with the arrangement of matter which instantiates it. Token-token identity, as we have already seen, would violate the principle of the irreducibility of PE. (It should be noted that we cannot appeal to supervenience to avoid the identity thesis because supervenience is not an ontological thesis.)

It might seem like the version of epiphenomenalism being developed here leaves too many loose ends, and forces us to accept too many counterintuitive conclusions, to warrant serious attention. However, the problems associated with epiphenomenalism are no greater than those faced by the contemporary materialist theories of mental causation. In opting for epiphenomenalism one merely trades one set of problems (of finding a place for PE in a world of physical causes) for the ontological, evolutionary, and epistemological problems that have been the subject of this chapter. Moreover, by choosing epiphenomenalism we have merely traded one set of counter-intuitive consequences for another; the only difference is that materialism has been around long enough for us to have become comfortable with its counter-intuitive elements.

To return briefly to the distinction between acceptance and belief, Popper makes the point that even Einstein never accepted (in a similar sense to the way that I have been using the term belief) the general theory of relativity and that:

[4] See. e.g. Block (1978)

'science may be regarded as a growing system of problems, rather than as a system of beliefs. And for a system of problems, the tentative acceptance of a theory or a conjecture means hardly more than that it is considered worthy of further criticism' (Popper 1994: 103). For the reasons that I have outlined above, I do believe that contemporary theories of mental causation are wrong, but I would not wish to say more of epiphenomenalism than that it is worthy of further criticism. If nothing else pushing the argument for epiphenomenalism as far as it will go will help us to work out what is wrong with epiphenomenalism. And, who knows, working out what is wrong with epiphenomenalism might even help furnish us with a final solution to the mind-body problem.

Part II: Epiphenomenalism and Social Theory

Chapter 6

Making Mind Matter More

My experience of wanting to know the time did not cause me to consult my watch, but neither was I merely following the stage directions in the situationalists' script. I, qua conscious subject, may only be a passive by-product of physical processes, but I am nevertheless a product of endogenous processes. There is, therefore, a very real sense in which I am a self-creation and more than just a gift of society. I spent part I of this thesis making mind matter less by arguing that the conscious mind is devoid of causal powers. Paradoxically, I now want to spend part II making mind matter more in social theorising and empirical research. That is to say, despite having claimed that the mind is devoid of causal powers I still want to argue that an understanding of a person's mental life is of key importance in providing a causal explanation of their behaviour. There will, of course, be certain limits on the type of mental events that, given epiphenomenalism, may be legitimately employed in casual explanations. This chapter will begin to address where these limits should be placed. Following these preliminary remarks the discussion will be expanded over the course of the chapters which follow (which deal with the self, rationality, and action respectively). As will become clear over the course of this chapter, although epiphenomenalism allows us to rule out phenomenal states as having any causal role, it does not provide much of an insight into how we should treat the content of mental states or accounts. As I noted in chapter 4, whether or not propositional attitudes turn out to be functionally discrete and semantically evaluable states (as defenders of PFP maintain) will have important implications for how we should go about causally explaining behaviour. This is an issue which is independent of the debate on mental causation versus epiphenomenalism. The version of epiphenomenalism that is being developed here holds that PE is epiphenomenal but does not deny that the content of mental states (conceived as neurally realised information) may be causally efficacious. Epiphenomenalism is, therefore, insensitive to whether PE supervenes on neural states that display the triad of characteristics identified by Ramsey et al (and discussed in chapter 4) or whether it supervenes on some property of a connectionist network (or, indeed, any other alternative).

Since the connectionism versus FP debate is still in its infancy it seems unwise for the social sciences to take sides at this stage – especially considering the fundamental methodological consequences that would result from their endorsing connectionism. Nevertheless it doesn't do any harm to be aware of the debate and start to consider its potential impact on the social sciences. More specifically, it might be helpful to be aware of which types of mental event causation can survive both epiphenomenalism and connectionism, if the latter turns out to be the correct theory of mind, and which it would be prudent to treat with a little more suspicion. Some of our conclusions regarding folk psychology and connectionism in chapter 4 are applicable here and will be expanded upon below. If epiphenomenalism and connectionism both turn out to be true then the consequences for our understanding of mental causation will be more far-reaching than if epiphenomenalism and the PFP turn out to be true. With this in mind I will tend to focus on the implications of epiphenomenalism plus connectionism in the remainder of this thesis, but will often examine the implications of both scenarios.

At this stage it might be worth reminding ourselves of the substance of our third principle, the principle of causal explanatory exclusion. This principle states that there can only be one complete and independent causal explanation of behaviour. The principle does not, however, rule out the possibility that there may be both physical and rationalising explanations of behaviour. Rather, the principle rules out the possibility that there may be a rationalising explanation and a physical explanation that are both complete and independent causal explanations. Given that our second principle, the principle of causal closure of the physical, arguably entails that there will always be a complete physical explanation for behaviour that cites only microphysical causal interaction, rationalising explanations must always clarify their relationship to 'in principle' physical explanations. In what follows we will assume that connectionism turns out to be true and provides something which comes close to the complete physical explanation of behaviour that our second principle states must exist. We can then begin to consider how different types of explanation relate to connectionism.

In chapters 1 and 2 we saw how PE is causally relevant in virtue of its supervenience on causally efficacious physical states. In chapter 4 it was further suggested that when propositional attitudes are consciously held it must be the case that the content of those propositional attitudes is physically realised. Moreover I argued that the content of consciously held propositional attitudes must be physically realised regardless of whether or not the brain turns out to be a connectionist network. Thus PE was again shown to be causally relevant because consciously held propositional attitudes are linked to action by a form of epiphenomenal causation. Unfortunately this causal relevance cannot be exploited for predictive or explanatory purposes. Since if epiphenomenalism turns out to be true phenomenal states have no effects whatsoever, the individual who is experiencing them cannot directly report them and they cannot be detected by any scientific instrument (if they could they would not be epiphenomenal since to be detectable is to have causal powers). The claim, therefore, that PE is causally relevant is to say the least somewhat misleading. What we really mean is that PE would be causally relevant if there were some means of telling when and what a person experiences.

Now the easy way out of this quandary would be to claim that actors' accounts of their actions, their citing of reasons, beliefs, hopes, and so on, covaries in a systematic and reliable way with their PEs. One could then say that if Jon says he wants to know the time, his account refers to his phenomenal time desire. We could then use the causal relevance of his phenomenal time desire to predict Jon's action of looking at his watch. This scenario is more or less what common sense would suggest and is how most academics treated accounts before the rise of social situationalism. Things, however, are not quite that simple. In order for Jon's speech act to refer to his phenomenal time desire, Jon would require access to his phenomenal time desire and this access is inconsistent with epiphenomenalism. How then are we to treat a person's accounts of their actions? Are we to assume that there is no relationship at all between an agent's accounts, their phenomenal experiences, and their actions? This would seem rather implausible. We can certainly assume that there is a determinate relationship between accounts and actions (though this, as we shall see in the final chapter, is not an assumption that is universally held). Such a relationship not only seems intuitively obvious but can be, and indeed is, demonstrated empirically on a daily basis since the very possibility of meaningful communication is dependent upon such a relationship.

Nevertheless, the conclusion that is imposed on us by accepting epiphenomenalism is that PE does not enter into the causal chain that culminates in actions and accounts. As we shall see, this does not mean that there is no relationship between PE and accounts, but it does mean that the relationship is not causal.

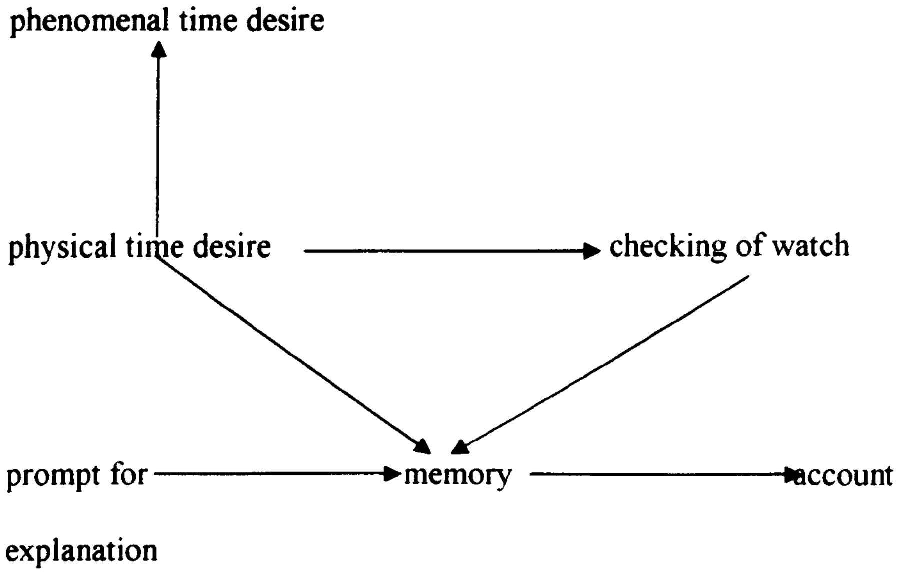


Fig. 6.1 The relationship between PE and accounts

Figure 6.1 illustrates the relationship between PE, actions and accounts from an epiphenomenalist perspective. In line with epiphenomenalism there can be no direct causal relationship between PE and either accounts or actions, but there is a direct causal relationship between the content of mental states (defined in terms of neurally realised information and functional physical states) and actions and accounts. This relationship allows us to explain why accounts are useful tools for the explanation of individuals' conduct. Of course the neurological story that underpins this model, were one available, would be considerably more complex.

For a start different brain regions are involved in the initiation of movement and linguistic ability so accounts and actions can not be directly linked by the same neural structures. In the above representation all these processes have been combined under the category of memory. Memory here is to be construed as the physical recall of neurally realised information. As I pointed out in chapter 4, how that information is stored, as a set of connection weights or as functionally discrete propositional attitudes and propositional memories, will have important implications for how we treat accounts. These implications will be developed in this chapter and those that follow. Although the details of the causal chain may elude us, if epiphenomenalism is true and there is a determinate relationship between actions and accounts, then it must be the case that actions and accounts are linked by a causal chain of neural events.

If the model outlined in figure 6.1 is along the right lines then connectionism may prove to be compatible with the standard FP and social scientific explanatory models for cases when a person cites propositional attitudes in their accounts. As I noted in chapter 4, however, even where propositional attitudes are realised by a connectionist network, and are causally efficacious in causing actions, it is unlikely that they will provide the complete cause. The reason is that the subvenient base for consciously held propositional attitudes (in the form of activation patterns) may form part of a larger neural state that is itself the cause of action.

Where connectionism poses a more serious threat is for cases where propositional attitudes are hypothesised to account for what appears to be purposeful behaviour (as when the existence of causally efficacious propositional attitudes are hypothesised to have caused observed behaviour or when non-conscious beliefs and desires are invoked in causal explanations). In such cases there is no means of establishing whether the behaviour in question was indeed caused by physical states that were the physical realisations of propositional attitudes (perhaps in the form of activation patterns) or if the behaviour was caused by physical processes that instantiated no mental content and caused no PEs.

Connectionism poses less of a threat to accounts that cite external factors. We can treat accounts that refer to external factors as referring more directly to the cause of behaviour than accounts which cite internal factors such as 'motives', 'intentions', 'reasons', and so on. This is because reasons that cite external factors refer to objective features of external reality (we do, of course, need to explain how knowledge of external reality is possible given epiphenomenalism, but we shall leave this task until our discussion of rationality in chapter 8). External factors can only causally influence behaviour, according to epiphenomenalism, by having an effect on the internal physical states that are themselves the cause of behaviour. When citing an external factor, however, it does not much matter whether the internal physical state it causes is an activation pattern or a functionally discrete and semantically evaluable state. As the following examples highlight, however, most actions are caused by both internal and external factors. These factors are then run together in accounts which means that in practice the distinction will rarely be clear cut. The following examples illustrate this point:

(i.) 'I chopped more firewood because it was a cold evening.' This account cites external reasons for action and is a sufficient causal explanation that does not require supplementation with internal causes. FPists typically want to extend such explanations to include beliefs, such as '*x* believed that the best way to stay warm would be to chop more wood for the fire,' and desires, like '*x* wanted to get warm.' Such additions are appropriate within the PFP explanatory framework when the aim is a precise causal explanation but are unnecessary for more general explanations since they are logically entailed by the more parsimonious account. Within an epiphenomenalist ontology based on connectionism, however, such additions are not only unnecessary but spurious. As we saw in chapter 4, if connectionism is true, the only physical states that possess content displaying propositional modularity are those that generate a PE with content. Thus unless it was in the conscious mind of the wood chopper that he desired warmth, and believed that the best way to get warm was by chopping more wood for the fire, then such states cannot form part of a causal explanation for the action. Nevertheless, temperature, as an objective feature of external reality, can still be invoked in causal explanations of wood chopping behaviour regardless of whether the PFP or the most radical version of connectionism (or any other possible explanation) turns out to be true.

(ii.) 'I don't want to tackle that overhang so I am going to avoid it by traversing to the left.' In this type of explanation the external world, in the form of an overhang, has had a causal impact on a climber's choice of route. The existence of the overhang is a partial explanation for the climber's decision to traverse but, assuming the overhang is within the climber's abilities, a complete causal explanation of the climber's decision would require reference to their internal states (such as fear or a desire to conserve energy).

(iii.) 'I took the philosophy course because I think that philosophy helps one to think clearly.' This account cites internal motivating factors and reasons that, if epiphenomenalism and connectionism are true, are at best only a partial cause of action.

By contrasting the way that these accounts might be treated if one applies the PFP model and an epiphenomenalist perspective we can see how the ontologies of these two perspectives inform the way that accounts are treated. The PFP, it will be remembered, assumes that mentalistic terms refer directly to semantically evaluable, functionally discrete, and causally efficacious states. For the PFP these characteristics combine to make all three explanations sufficient causal explanations. In the third case, for example, the belief that 'philosophy helps one to think clearly' would be seen by PFP as a discrete state that, combined with other states, such as a desire to think clearly, directly cause the action of signing up for a philosophy course. An epiphenomenalist ontology in contrast, particularly one with connectionist leanings, does not view the belief that philosophy helps one to think clearly as an internal state with the triad of characteristics attributed to it by PFPists. As such the account cannot be seen as providing a sufficient causal explanation (and nor does it cite a necessary condition for the student to sign up for the course). Nevertheless, the account can still be used to predict behaviour since if a person says that they believe philosophy is a discipline that helps one to think clearly, and they claim to view thinking clearly as desirable, then this provides good grounds for predicting that they will sign up for a philosophy course. When using accounts in this manner we are relying on the empirical observation that people tend to do what they say they are going to do, want to do, and so on. As this suggests, if epiphenomenalism is true then prediction and causal explanation cannot be treated as two sides of the same ontological and methodological coin. Epiphenomenalism plus connectionism allows for the possibility of accurate prediction based on a flawed PFP style ontology. That is to say, by adopting the intentional stance, the language and ontology of FP might be usefully employed as a methodological resource for the prediction of behaviour even if the terms employed do not refer to causally efficacious states. This analysis of accounts still leaves unanswered the question of how one should go about causally explaining actions and what role accounts (especially accounts of the third type) should play in casual explanations. I will have more to say about this topic in chapter 9 but at this point I will say that I think that the answer will be largely an empirical one that is beyond the scope of this thesis to answer.

Turning now to the remainder of this thesis, I have already voiced certain reservations about both the ontology and the methodology of contemporary sociology. In chapters 7 and 8, which deal with the self and rationality respectively, I will review what I consider to be some of the principal failings of contemporary sociology and offer some preliminary suggestions as to how the ontology of epiphenomenalism can be used to address some of these failings. Both chapters are essentially a defence of human subjectivity, and the role that I believe subjective states should play in causal explanations, against the contemporary decentring /desubjectivising of the subject. In chapter 7 I will review the attempts made by sociologists, and Mead in particular, to undermine the autonomy of the self and its biological origins. Here I will argue along similar lines to Margaret Archer who claims that sociological imperialists have conflated the universal sense of self with the (cultural) concept of self. I will suggest that such approaches conflate the self with self-identity and in so doing portray man as a one dimensional linguistic construct that is incapable of true action.

My defence of a universal sense of self with origins in our biology and interaction with the external world does differ markedly from traditional concepts of the self. I will argue that the self is constituted by the PE of subjectivity. This self, being constituted by PE, is an epiphenomenal self that is generated by physical events in the brain in exactly the same way as PE generally. Conceived in this way the self is denied both a will capable of initiating action and its traditional role as the seat of reason. The latter characteristic has some profound implications for our understanding of rationality and these implications will be examined in chapter 8. Specifically we will have to confront the worry that if the self is not equipped to weigh up evidence and to make judgements based on that evidence then there can be no good grounds for accepting an argument as true or false. This not only has the potential to undermine the truth claims made in this thesis but to signal a complete collapse into relativism. Evolutionary epistemology can be usefully employed in order to shore up the foundations of rationality that the denial of an autonomous self undermines. Karl Popper's evolutionary epistemology in particular, and his argument that there can be knowledge without a knowing subject, will provide much of the groundwork. Popper's work provides the first step in an argument designed to show that physical systems are capable of generating knowledge and evaluating truth claims without the intervention of an irreducible self.

In chapter 9 I continue the theme, first outlined in chapter 7, that the ontological stance of contemporary sociologists makes human beings incapable of true action. I will begin by outlining Campbell's view that action has been written out of contemporary sociology and has been replaced by an impoverished concept of social action. This will be followed by the claim that sociology should return to a Weberian style interpretative method. To return to the example that opened this chapter, the act of looking at one's watch is more likely to be explained by social situationalists as signalling to those present that, for example, the actor is not loitering around waiting for an opportunity to abduct one of their children, but is instead engaged in the perfectly legitimate activity of waiting for a friend who happens to be late. I would not wish to deny that such explanations can, on occasion, be true. What I will argue (in chapter 9), however, is that one cannot and should not presume to understand an individual's motives for performing any given action without taking the trouble to ask them. As we shall see it is a feature of the dominant paradigm within sociology (social situationalism) that the actor's subjective states and intrapersonal processes are largely ignored within both social theory and empirical research. The focus of contemporary sociology has now moved from the individual as the author of both their private and social lives (within, of course, certain constraints) to the individual as an actor reading from a collectively written script. The odd thing about this collaborative project is that the script is viewed as emerging out of interaction rather than being the culmination of the efforts of each participating individual. The upshot of all this is that there is now almost no interest in anything internal, or 'in the head' of the individual. Sociology has thus constructed for itself an ontology and a methodology within which the simplest of actions, such as consulting one's watch, become impossible to explain as anything other than communicative acts. It is in this sense that I want to make mind matter more by highlighting the urgent need to make what's in the head central to sociological explanations.

Chapter 7

The Brain and its Self

In a statement that captures most people's intuitions regarding the relationship between selves and brains, Popper wrote in *The Self and Its Brain*:

> ... the brain is owned by the self, rather than the other way round. The self is almost always active. The activity of selves is, I suggest, the only genuine activity we know.[1] The active psycho-physical self is the active programmer of the brain (which is the computer), it is the executant whose instrument is the brain. The mind is, as Plato said, the pilot. It is not, as David Hume and William James suggested, the sum total, or the bundle, or the stream of its experiences: this suggests passivity. (Popper and Eccles 1977: 120))

Popper, dualists generally, and a good many materialist theorists need an active self to complete their various theories of mind. The self is traditionally deemed to perform a number of functions: (1) the self initiates voluntary action; (2) it is the subject of conscious experience, thus unifying what many believe would otherwise be a disjunctive 'bundle' of experiences; (3) it is able to entertain first-person thoughts, which is crucial for the self-other and subject-object distinctions that so preoccupy philosophers; (4) and is capable of self-knowledge and introspection. The latter feature is often held up as vital for the ability of a person to recognise that past and present experiences are theirs, and to conceptualise their life trajectories (ending of course in death) enabling them to organise their desires, aspirations, and goals in the context of decision making.

Epiphenomenalism, of course, requires no such self. For epiphenomenalism: (1) there are no voluntary actions – only causal processes; (2) there can be no subject of conscious experiences (at least not of the traditional Cartesian kind) since this would imply that, at the mental level at least, phenomenal states have causal

[1] By 'genuine activity' Popper is presumably referring to agent causation (in the sense of genuine libertarian free will). rather than the standard relation of cause and effect which would not constitute action for Popper.

powers. That is to say, if we draw the traditional distinction between the subject of conscious experiences and the conscious experiences of which a subject is aware, we accord to phenomenal states the causal power of being observable. Moreover, we accord to the self the causal power of being able to introspect. Neither power is consistent with epiphenomenalism which denies both mental-physical and mental-mental interaction. This point also extends to both (3) the ability to entertain first-person thoughts and (4) to have certain kinds of first-person knowledge and introspective abilities. For epiphenomenalism both these features are functions of physical processes not of the self.

It is a good thing that epiphenomenalism does not need an active, unifying self, since as the above makes clear such a self is unavailable to the epiphenomenalist theorist. Nevertheless the dualism expressed by Popper (between the self and body) is one that we can endorse. Throughout this thesis the *sui generis* real nature of PE has been emphasised. This emphasis extends to all PE, including the PE of selfhood, thus guaranteeing the existence of the self – albeit a self devoid of causal powers. We can also agree with Popper's statement if we read it as a phenomenological description of the experience of selfhood. As an ontological statement, however, it is the complete antithesis of the theory to be presented here. In part I we spent a considerable amount of time discussing and developing the idea that PE supervenes on physical events in the brain. The supervenience of the mental on the physical, it will be remembered, entails that all phenomenal states are wholly dependent on and determined by physical events in the brain. Since these dual characteristics of dependency and determinacy appear to leave no room for mental causation we concluded that PE must be epiphenomenal. This leaves us with a fundamentally dualist ontology –

though not of the traditional kind – and begs the question, where does the self fit into this ontology? We must choose between conceiving of the self as some physical property of the brain, thereby guaranteeing it causal efficacy, or as something irreducibly mental and hence epiphenomenal? As the title of this chapter suggests, I will argue in favour of the latter option. Although epiphenomenalists can agree with Popper that the self is ontologically distinct from the brain, what an epiphenomenalist perspective forces us to deny is that the direction of causation runs from self to brain. Where Popper views the self as the active programmer of the brain, we are forced to view the self as its passive creation. The passivity of the self, together with its dependency on the brain, is what warrants the inversion of Popper's dictum which now reads *the brain and its self.*

Identifying the self with some irreducible and epiphenomenal mental property does present us with some serious methodological problems. Throughout this thesis we have been taking certain liberties with the ontological implications of epiphenomenalism. If PE really is epiphenomenal we ought not to be able to discuss its existence, let alone use our 'knowledge' about PE as the basis for further ontological claims. Nowhere is this more apparent than when discussing the self. The self is something which exists for us in PE and which we know only through PE. Introspection is thus one of the primary methodological tools for the investigation of the self. If we were to hold steadfastly to implications of accepting epiphenomenalism, however, we would not only deny ourselves this method of investigation but we would have to remain mute on the whole topic. Any epiphenomenalist that attempted to use their own experience of selfhood as the basis for a discussion of the self would have to accept that their theories and conjectures, as well as the words they type on their computers, are the result of neuronal processes that have no connection with their subject matter. This point also highlights a problem for rationality that will be discussed in the next chapter; if arguments, judgements, theories and conjectures, are just a product of blind physical forces, what reason could there be for our accepting them as true. If we are to avoid epiphenomenalism from leading to complete scepticism and nihilism, however, then it seems unavoidable that we will have to take further liberties and hope that some future justification will be found for our indiscretions.

The principal ontological liberty I propose to take is to begin our discussion by outlining the experience of selfhood. We will then test the characteristics belonging to the experiential self against the conclusions drawn in part I. As ever we must treat the experience as its own irreducible reality and not attempt to explain it away in some vain attempt to construct a reductionist account of the self. Such attempts, or so I argued in part I, are doomed to failure since they end up either eliminating the object of reduction or describing its cause while leaving the object of reduction untouched. The latter form is typical of the biological sciences. Neurology, for example, is beginning to yield some detailed accounts of the neurobiological processes implicated in conscious experience, if PE is irreducible, however, such accounts relate only to the causes of conscious experiences rather than the experiences themselves. Thus far we have treated, following Searle's lead, PE as its own irreducible reality. This approach suggests that phenomenology should be treated as a branch of ontology, which if we remain committed to our first principle regarding the irreducibility of PE must be the correct way to conceive of the relationship.

When discussing consciousness there is always a temptation to ask what is *x* (where *x* is some PE). Where PE is concerned, however, the answer will always be either a tautology or a causal reduction. If one asks, 'What is the pain in my foot?' for example, one can answer in one of two ways. Either (a) it is a subjective experience of an unpleasant sensation that I experience as being spatially located at the end of my leg (tautology), or (b) it is nociceptive specific neuronal activity occurring in the somatosensory cortex (causal reduction). We have already argued at length that this style of reduction is unacceptable since it fails to answer the question of what the pain is (an ontological reduction) and instead tells us what causes the pain (a causal reduction).

Taking methodological liberties with epiphenomenalism is essential if we are to avoid such reductionist or behaviourist accounts (which would seem to be the only other legitimate alternative). Being dutiful epiphenomenalists, and refusing to countenance the use of PE as a source of information about the self, would necessitate our adopting a behaviourist methodology that would inevitably result in our account of the self missing out the defining characteristic of selfhood as that which experiences the world. This seems to have been the fate of, amongst others, Goffman's dramaturgical analysis of the self. Theories such as Goffman's, which view the self as a product of social interaction and locate the self in the performance of actions, may be more kosher from an epiphenomenalist perspective than those that use introspection as a primary methodological tool, but they end up defining the self out of existence by excluding subjectivity from their definitions. If the 'very structure of the self' is identified with the arrangement of the presentation of self and if the possessor of a self 'merely provides the peg upon which something of collaborative manufacture will be hung for a time' (Goffman 1971: 245) then the experience of selfhood is implicitly nullified. Treating the appearance as the reality means accepting that whatever external accounts, such as Goffman's, claim to be about, they are not about the self. That said, however, we must resist the temptation to take the experience of selfhood at face value and draw ontological conclusions from the experience alone (as in Descartes' Cogito).

This chapter is going to be a largely negative discussion of the self. I will have a lot to say about what the self is not, what conditions are not necessary for the emergence of the self, and what properties and powers are not possessed by the self. These negative conclusions will provide the basis for the claim that the universal sense of self is not a legitimate area of study for the social sciences. The universal sense of self, it will be argued, develops quite naturally out of our ongoing interaction with the world and does not require society or its resources as a necessary condition for its emergence.

The experiential self

Gilbert Ryle nicely captures the puzzlement that confronts all of us when we reflect on the concept of the self:

> When a child, like Kim, having no theoretical commitments or equipment, first asks himself, 'Who or What am I?' he does not ask it from a desire to know his own surname, age, sex, nationality, or position in the form. He knows all his ordinary personalia. He feels that there is something else in the background for which his 'I' stands, a something which has still to be described after all his ordinary personalia have been listed. He also feels, very vaguely, that whatever it is that his 'I' stands for, it is something very important and quite unique, unique in the sense that neither it nor anything like it, belongs to anyone else. There *could* only be one of it. (Ryle 1949: 178)

Such is the sense of mystery that we have all experienced at some time or another. For those of us who continue to ask the question in an academic context our 'theoretical commitments and equipment' does little to lessen the sense of mystery. One thing is certain however, that whatever the self is (if it is indeed a thing) it is something whose defining characteristic is that it is conscious. We may not be able to pin down the self but if it is to be found anywhere it is to be found in our conscious experience and it is with this experiential self that we will begin.

We have the experience of there being an *x* that:

1. is conscious or potentially conscious
2. is indivisible
3. is unique
4. is elusive
5. has language as its prime (almost exclusive) medium of thought
6. is the locus of perception
7. persists through time
8. is capable of engaging in introspection
9. makes decisions based on reasons
10. acts on reasons and decisions to cause actions
11. is responsible for its actions

Despite having criticised reductionist and external accounts of the self for failing to treat the experience of selfhood as the reality, as epiphenomenalists we cannot treat all the properties and powers that appear on the above list as genuine ontological traits. The trouble is that where the self is concerned we are not only dealing with sensations and perceptions but also with cognitive judgements. If we consider our list of the features that go to make up our experience of the self, points 1 to 7 describe experiences which we have to accept as the reality if we are to continue treating phenomenology as a branch of ontology. We would, of course, have to be clear about what we mean by such things as uniqueness or elusiveness, but each of these elements has a fairly straightforward interpretation that relates solely to the experience and has no ontological implications beyond treating the experience as the reality. Thus, for example, we might define uniqueness by noting that every person has a unique perspective on the world and a unique set of experiences, and we might define elusiveness as a feature of subjectivity (we will consider this feature in detail below) rather than as pointing to the existence of a thing or substance for which the 'I' stands. The last four points, however, though they too describe the phenomenology of selfhood, also make ontological claims about the powers of the self. Here, though we have to accept that the experience is partly constitutive of the self, we do not have to accept that selves have the causal powers they experience having. So, for example, although we have to accept that the *experience* of free will (implied by points 9, 10 and 11) is an ineliminable part of the experience of selfhood, we do not have to accept that selves really are in control of their bodies.

Nevertheless, we must still confront the fact that during the lived experience of our self and body we view our bodies, including our brains, as something which belong to us and that, most of the time, we control. Although very few of our movements are accompanied by the experience of being consciously 'willed' this does not result in our feeling that we do not control our bodies. Habitual or routine actions are typically in accord with our desires, projects, goals, etc., and as is often noted with regard to actions that have become habitual, though we do not have the experience of consciously executing the actions we often have the experience of initiating them. Thus, although we often put our bodies on autopilot when walking to the corner shop or driving to the supermarket, we nevertheless believe that the initiations of these actions are consciously willed.

The control we believe we have over our bodies is central to the feeling of psychophysical dualism to which Popper points. Psychophysical dualism is also evident whenever we think reflexively about sensations, especially when those sensations are unwanted. Vrancken (1989) makes a similar point arguing that rationalising pain induces the experience of psychophysical dualism by sundering the subject 'I' from the pain as object. Our experience is, therefore, of an agency that controls the body but is not identical with the body or any of its parts. This must be the case since if we were to experience our selves as identical to our bodies we could not have the experience of our bodies acting without us (as in reflex actions) or in spite of us (as when our bodies become paralysed with fear) and nor could we experience the psychophysical dualism to which Vrancken points. There is, of course, an intimate connection between our *experience* of the self and the body but the connection is, as Popper highlights, one of ownership not identity. There can be no escaping the fact that we view our *physical* bodies as being under our *mental* control. If epiphenomenalism is true then we have to accept that this psychophysical dualism points to a genuine ontological division between the self and the body. The self, therefore, cannot be identified with the body (or any part of it).

So where does this leave us? We have rejected reductionist accounts of the self, those that locate the self in the performance of actions, and those that attempt to identify the self with the body. We have also suggested that only points 1 through 7 on the above list point to any genuine ontological traits that we may use to define the self. Even points 1 through 7, however, will be pared down before we will be left with the essence of selfhood. To help us pare down this list I want to focus on the quality of elusiveness.

His quarry was the hunter

The self, we have concluded, is to be found in PE. As such it might be expected that epiphenomenalism would favour a Humean concept of the self where the self is identified with the sum total of one's occurrent PEs. Such an approach would be consistent with the dependency of the self on the brain, it would locate the self solely in PE, and would be consistent with our conclusion that PE is epiphenomenal. A Humean version of the self, however, despite originating as an antidote to problem of elusiveness, does not do justice to the experience. A Humean version of the self explains elusiveness away as an illusion rather than embracing it as part of the reality of selfhood. Hume, it will be remembered, famously proposed the 'bundle' theory of the self because he was never able to catch himself without some perception, and was never able to observe anything but the perception. The trouble is that whenever we ask ourselves questions such as: Is the sound of bird song that I hear through an open window part of me? We are compelled to say no. The sound of bird song is something I am aware of but it is not part of me. Perhaps then we should say that those experiences that derive from proprioception are what constitute the self. This doesn't seem quite right either though. I do not count the bitter aftertaste of coffee, the awareness of the position of my limbs, or feelings of hunger to be part of my self. Going 'deeper' we might think that our beliefs, likes and dislikes, personality, and the rest are what constitute the self. Once again, however, the self eludes us; I do not count my desire for another coffee, my interest in the mind-body problem or my belief in epiphenomenalism as constituting part of my self. I might say that my interest in the mind-body problem is part of what makes me *me* but this is not the same as saying that my interest in the mind-body problem partly constitutes me. Ryle's explanation for the elusiveness of the 'I' is that to reflect on the nature of one's self requires a higher order act and this '... higher order action cannot be the action upon which it is performed' (Ryle 1949: 186). Ryle sums this up with the evocative phrase, which I have taken as the title to this section, 'His quarry was the hunter' (ibid. 189). Nevertheless, Ryle thinks that he could exhaustively describe his self in the past tense. Indeed Ryle even believes that he could exhaustively describe your past or present self. This seems doubtful, however, since if we reflect on last year's self and ask the same questions we asked of the present self we would get the same answers (with my desire for another coffee now satisfied I do not consider that my self of five minutes ago was constituted, in part, by a desire for another coffee). Moreover, Ryle's explanation for the elusiveness of the self is not consistent with the epiphenomenalist perspective being developed here. When Ryle notes that a 'higher order action cannot be the action upon which it is performed' he is making the now familiar claim that a mental event cannot introspect itself. Rather, according to Ryle, a second mental event is required to introspect other synchronous mental events. Epiphenomenalism, of course, denies the possibility of this type of introspection and so must reject Ryle's explanation for the elusiveness of the self.

Phenomenologically it does appear to us that we are able to introspect our mental states. Psychophysical dualism is an endemic feature of conscious experience that extends beyond the objectification of our bodies to include the potential objectification of all phenomenal states. Though the experience may be illusory we seem able to make every perception, sensation, verbalised thought, emotion, or indeed any PE, the objects of our attention. Illusory or not, the experience of being able to objectify PEs means that we cannot identify ourselves with those experiences. Rather, we believe ourselves to be the subject of those experiences. As has been pointed out may times before, when Hume went in search of his self and was never able to find anything but some perception or sensation, we have to ask ourselves who was doing the searching. Hume could not find the self that he was looking for because, as Ryle puts it, his quarry was the hunter. Where Ryle went astray, however, was in imagining that the hunter is only elusive in the present tense and that a description of the hunter's experiences in the past tense is a description of a person's past self. The trouble is, as I noted above, any description of my past experiences, no matter how complete, will always fail to describe my past self because I will always believe that I am (or was) the subject of the experiences being described. If this is right then it seems that subjectivity must be the key to the self. Though it is of course true that we are always subjectively aware of something and can never experience a state of pure subjectivity, it seems that subjectivity is a distinct and irreducible component of every experience, and it is this component of experience that is the self.

Origins of the self

There is a strong tendency within the social sciences to assume that the self is either a linguistic construct or somehow dependent on language. The first point to note is that by making the self dependent on language such theorists are forced to deny selfhood to organisms (including humans) that are not language users. Such a denial seems, to say the least, rather anthropocentric. Beings without language are still capable of experiencing psychophysical dualism, they are able to objectify their bodies, sensations and emotions, and they are capable of engaging in practical action. Were they unable to draw a distinction between themselves as the subject of experiences, and the experiences of which they are aware, then beings without language would be unable to respond to their environment or to perform even simple actions such as withdrawing from painful stimuli. Even pain avoidance behaviour presupposes that the organism is capable of recognising that the pain is theirs but that it is not an intrinsic part of their self. That is to say, in order to act they have to be capable of making a subject-object distinction (where the pain is confronted as object). Once an organism is capable of making such distinctions they are already in possession of a rudimentary sense of self.

One way social scientists and philosophers have made the development of the self dependent on language is by associating the self with narrative. Rorty, for example, views selfhood as a process of self-creation whereby we come to terms with the 'blind impress' which chance has given us by redescribing the contingencies of the past in our own terms (Rorty 1989: 23-43). Rorty's association of selfhood with narrative, and its concomitant dependency on language, ignores what I have argued is the most fundamental aspect of selfhood – human subjectivity. Moreover, it is unclear how a person is able to redescribe the contingencies of their own past unless they are already in possession of a sense of self. Just whose past are they attempting to redescribe? If it weren't *my* past, a past that *I* remember, then I would have no interest in its redescription. Rorty's account may be an accurate description of how a few postmodern intellectuals view the reflexive construction of their own self-identity, but it has nothing to do with the universal sense with which we are all blessed.

Where Rorty views narrative structure as constitutive of the self, others view language as a necessary precondition for the development of the self. Above all others it is probably George Herbert Mead who is responsible for this belief. Central to Mead's philosophy of mind is the claim that the existence of the mind is dependent upon the 'internalized conversation of gestures' (Mead 1934: 156). According to Mead (ibid. 47), we learn the meaning of words and gestures, and are conscious of the meaning of our own utterances and gestures, only in so far as we take the attitude of the other towards them.

> Gestures become significant symbols when they implicitly arouse in an individual making them the same responses which they explicitly arouse, or are supposed to arouse, in other individuals, the individuals to whom they are addressed; and in all conversations of gestures within the social process, whether external (between different individuals) or internal (between a given individual and himself), the individual's consciousness of the content and flow of meaning involved depends on his thus taking the attitude of the other toward his own gestures. (ibid. 47)

This internalised conversation of gestures is what, for Mead, constitutes thought and the life of the mind. Thought is thus dependent, both ontogenetically and phylogenetically, upon the pre-existence of language and society. That is to say, the potential for conscious thought is not, for Mead, an innate potential inherent in every member of the species and nor could consciousness develop in any given individual unless they were socialised in a language using community. Similarly, the development of the self is, for Mead, dependent on taking the attitude of others toward oneself. Here Mead argues that there are two stages in the development of the self, both involving the internalisation of the attitudes of others but at neither stage does subjectivity play a primary role.

> At the first of these stages, the individual's self is constituted simply by an organization of the particular attitudes of other individuals toward himself and toward one another in the specific social acts in which he participates with them. But at the second stage in the development of the individual's self that self is constituted not only by an organization of these particular individual attitudes, but also by an organization of the generalised other or the social group as a whole to which he belongs. (ibid. 158)

For Mead this process is logically and temporally prior to the development of self-consciousness. In other words, knowledge of how others view us must occur prior to our having any knowledge of ourselves. What is not clear is how self-consciousness could be constructed by organising the attitudes of others without there already being some reference point for those attitudes. In order for me to be able to adopt the attitudes of particular others toward myself, which for Mead constitutes the primary phase of the development of the self, I have to be able to differentiate myself from others and be in possession of a basic concept of selfhood in order to know that others' attitudes are about me. To do so I have to be able to draw on a fairly sophisticated folk psychology, though of course I need not be able to articulate the content of that folk psychology. To be able to internalise the attitude of a particular other one requires a working knowledge of propositional attitudes; I have to know that the other person is a conscious subject who has certain beliefs and attitudes about me. In short, *I have to know that I am an I in order to know that I am a you to you and that your attitudes are directed toward me.* By this I do not mean to suggest that one has to master the use of personal pronouns before one can internalise the attitudes of others. Even animals such as dogs, which possess no linguistic ability, are adept at making these distinctions and are perfectly capable of understanding that, for example, their owner is angry at them for raiding the bins.

What has led Mead and others astray is the conflation of self-identity with the self.[2] If, by self-identity, we mean a narrative about oneself, and if we read Mead's account as a theory concerning the construction of that narrative, then

[2] Margaret Archer argues along similar lines when arguing that sociological imperialists, like Mead, have a tendency to absorb the sense onto the concept (see Archer 2000: 125).

there is far less about Mead's work with which to quibble. Since we require the resources of a society (its language, definitions of social roles, norms, values, and so on) in order to construct a narrative about ourselves, the pre-existence of society is a necessary condition for the development of self-identity. Moreover, much of the content of the narrative we tell about ourselves is dependent upon our having internalised the attitudes of others; at least to the extent that we know what it is to be a good father, a dutiful son, etc. Although it is always open to us to question, and possibly reject, the opinions others have about us, it is nevertheless the case that others' opinions are of great importance to us (indeed the preoccupation with the opinion of others pays the wages of the growing army of therapists, psychologists and counsellors). Moreover, though man is certainly capable of self-deception, the deception is difficult to maintain in the absence of external corroboration.

The self and self-identity are ontologically distinct and must be kept analytically distinct since their conflation is not only a denigration of humanity but represents sociological imperialism of he most insidious kind. As Ryle comments in the above quote, when a person asks themselves who or what they are, they are not interested in listing their surname, age, sex, or position in the form. Neither are they interested in listing their occupation, sexual orientation, religious affiliation, or any of their desirable qualities (such as being a dutiful son). Strip away the narrative and we are still left with a subject who stands in certain relations to his body and to the external world. It is this subject that is the something in the background for which the 'I' stands and not those attributes of a person that can be listed or those mental events that can be experienced as the objects of subjective awareness. The fact that we can take a step back from our self-identity and treat it as a 'reflexive project' presupposes that self-identity is not synonymous with the self. There is no vantage point to which one could withdraw and treat the self as a reflexive project because the self is that vantage point. It is for this reason that Giddens is wrong to treat the self as a 'reflexive project' (Giddens 1991) and Rorty is wrong to view the self as a self-created narrative in which we redescribe the contingencies of the past in our own terms. Of both accounts we can ask fundamentally the same question, who is treating the self as a reflexive project and who is redescribing the contingencies of *their* past?

Despite arguing from a different ontological perspective Margaret Archer's discussion of the primacy of practice in the development of the universal sense of self is remarkably compatible with epiphenomenalism. In asserting the primacy of practice Archer is defending the idea that our practical engagement with the world has temporal and logical priority over language in the development of the sense of self. Archer argues that socialisation is dependent on:

> (a) the fact that each and every member has already realised one potential of their species-being, namely to make the primary distinction between the self and otherness. on which learning the subsequent distinctions between social and non-social depend. Socialisation (b) requires human beings with performative capacity and memory in order that they are the kind of beings whose repertoire can be socially extended to incorporate such activities as handling a spoon, becoming toilet trained or leaning to speak. Finally, (c) the very possibility of communication, whether gestural or verbal, is ultimately dependent upon beings who are already obedient to the law of non-contradiction, otherwise no verbal information can be conveyed, including natural language itself, as distinct from mimetic babbling. (Archer 2000: 126)

Taking these elements in turn, for epiphenomenalism (a) learning the self-other/subject-object distinction can not be matter of the self learning to distinguish itself from its environment. Having stressed the importance of subjectivity as the defining characteristic of the self this would amount to a form of panpsychism whereby subjectivity shrinks from a starting point that encompasses the body and its environment to a perspective within the body. Such a position is obviously untenable given that perception is necessarily embodied. Rather, when we talk about a person learning the self-other/subject-object distinction we mean that the sense of self develops in parallel with non-conscious (neural) development that enables the organism to make the distinctions. Notice that it is the organism that learns to make this distinction not the self. Given that we have defined the self in terms of subjectivity, 'learning' the self-other distinction is necessarily prior to, or at least synchronous with, the development of one's sense of self. It is also important to stress that the completion of the neural architecture that enables the organism to make the subject-object split is not identical to the development of the sense of self. The magic ingredient of subjectivity is, from an epiphenomenalist perspective, still missing. This subjectivity emerges (in part) out of this architecture but is not identical with it. Thus when Archer asserts the primacy of practice in the development of the self, epiphenomenalism inserts a second link in the causal chain. What for Archer is the origin of the sense of self is, for epiphenomenalism, the origin of the physically necessary (at least in humans) conditions for the emergence of a sense of self. In other words, where Archer views practical action as contributing to the development of the sense of self, we view the same action as contributing to the development of the subvenient base that will eventually give rise to the supervenient and epiphenomenal property of a sense of self. In the final analysis epiphenomenalism does not have the resources to explain the origins of the sense of self. As epiphenomenalists we have to wait for the natural sciences to discover how the brain generates consciousness. As epiphenomenalists all we can hope to achieve is an understanding of the environmental conditions which are physically necessary for the development of the neural structures that will ultimately generate an epiphenomenal sense of self. Perhaps the following diagram will make this clear:

body-world interaction → contributes to the development of the neural architecture which causes → an epiphenomenal sense of self

Fig. 7.1 The origin of the sense of self

Epiphenomenalism may be able to say something about the first link in this chain, but must remain mute on the question of how neural processes generate a sense of self. Critical realists, identity theorists, emergentists, and so on, are equally in the dark about the neural structures that relate to the sense of self. However, because they view the self as either directly efficacious, or efficacious in virtue of its identity with its subvenient base, this unexplained element is not viewed as a distinct link in the causal chain. Rather, they view their accounts as complete causal explanations at a higher level of analysis than that provided by the natural sciences (which are left to fill in the details).

Archer's second argument for the primacy of practice, that socialisation requires beings with performative capacity and memory, is fairly straightforward in relation to the development of the sense of self. Performative capacity is necessary for practical action and must, therefore, develop before a capacity for meaningful social action. Since even feral children and animals have a performative capacity, we can safely assume that it is not dependent on society or its resources. Memory too, in a broad sense, is obviously not dependent on linguistic ability (dogs can remember the location of their favourite toy and horses can remember which gardens are home to ferocious dogs). Of more interest is Archer's claim that eidetic (visual) and procedural memories are more important in the development of a sense of self and its continuity than declarative memory. I have argued that the self is constituted by the experience of subjectivity that accompanies conscious experience. We have also noted the truism that one cannot experience subjectivity in a pure form. To be subjectively aware is to be subjectively aware of something. Thus eidetic memory, which Archer notes (citing Rose 1992) develops before declarative memory, provides a source of relived experience that links present subjectivity to past experiences enabling a person to maintain a continuous sense of self. It seems unlikely that declarative memory could perform this function since it lacks the temporal dimension of eidetic memories. That is to say, eidetic memories are time-indexed to the self that *I was*, whereas declarative memories lack this relived experiential component. Thus, because declarative memories are experientially atemporal, they cannot provide the basis for the continuity of the self.

The role of practical action as the source of our reasoning abilities is Archer's third argument in favour of the primacy of practice. Not only does it help to take back one more of mans' abilities from out of the jaws of the social imperialists, but it provides a good basis for our discussion of rationality in the next chapter. In the following chapter we will have to account for mans' ability to reason in the absence of a self capable of weighing up evidence and making judgements based on good reasons. If we can show that practical action is sufficient to enable a child to learn the rules of non-contradiction and identity in a way that is consistent with epiphenomenalism then we will be one step closer to achieving this task. Embodied practice is held by Archer to be necessary for the development of our powers of thought since these powers require a tacit understanding of the central tenets of logical reasoning: the laws of non-contradiction and identity. Archer argues that practical action is essential for our understanding both principles. It should be stressed that developing an understanding of either of these principles does not require that we are able to discursively formulate them or, indeed, conceptualise them in any overt way. Rather these principles form part of the background assumptions that enable effective interaction with the world and provide the necessary pre-conditions for rational thought (both discursive and non-discursive). The following example helps to clarify this point:

> ...the child lying prone in its cot, who moves away from the uncomfortable pressure of the cot bars is displaying mastery of both principles [of the 'object concept' necessary for identity and of 'conservation'/intransitivity necessary for both identity and non-contradiction]: she believes that the pressures come from without and that she can detach her body from the source of discomfort by rolling over and, moreover, that the bars are stationary in nature and will not come after her. (Archer 2000: 146)

The child in this example (and we may note that the example could just as easily featured an animal) is displaying what we might term an embodied rationality. For our purposes it is important to note that the child is not yet in possession of a developed sense of self, rather, they are in the process of developing a sense of self. Moreover, the child has not yet mastered the use of language. We have, therefore, an example of practical reasoning which is not dependent upon the existence of an irreducible self (thus helping to establish that the self is unnecessary for the operation of rationality) or of language, which demonstrates that the foundations of rational thought do not depend on society or its resources.

Continuity of the self

One important issue that we have yet to address is the continuity of the self. One of the most commonly cited features of the self is that it persists through time. At first sight this appears indubitable, it is certainly an indubitable fact of experience that I perceive myself to be the same person today that I have always been (at least for as long as I can remember I have been me). Moreover, I have no reason to doubt that barring the misfortune of, for example, total amnesia, severe brain damage or mental illness, I will remain essentially the same person for as long as I remain conscious. Having defended the idea that the self is the subjectivity that we always feel lurking behind experience, we have implicitly denied that the continuity of the self can be accounted for in terms of the continuity of a thing or substance. In previous chapters I argued that all PE supervenes on the synchronous internal physical states of the organism. This dependency of one's sense of self on synchronous internal physical states entails that the self must be, like all PE that persists through time (for however short a duration), literally created and recreated by physical events for the duration of a person's conscious life. A good parallel to draw here is between the sense of self and an eternal flame. Both are created and recreated by their physical cause (neural events in the case of the self and, for example, a gas burner in the case of an eternal flame). Now the question we must pose is in what sense is an eternal flame the same flame now as it was when it was ignited as a memorial to the dead in, say, the siege of Leningrad and what makes me the same person now as I was ten years (or even ten seconds) ago. In neither case does strict numerical identity seem to do the job. Gas burners may be replaced and repaired and the physical structure of the brain is continuously changing. Similarly, the volume, heat, colour, and physical composition of an eternal flame or the qualitative experiences of a person are never the same from one moment to the next. Since quantitative identity does not look like it is up to the job of securing the continuity of the self, we are forced to look towards a qualitative theory of identity. The literature on this topic is voluminous and we could easily devote a great deal of space reviewing the alternative theories and searching for counter examples. However, there seems to be no consensus amongst philosophers regarding suitable criteria for identity and little prospect of a consensus being reached. Such an endeavour would, therefore, be unlikely to reward our efforts. A more fruitful approach might be to set aside these issues and trust that our experience of being the same person over time is veridical. Having done so we may then seek to explain why this is the case. The search for this explanation could, if we are not careful, degenerate into the search for a criterion of identity if we approach the question by asking what makes it true that my experience of the continuity of self is veridical. A better way to approach the question is by looking to the experience of continuity and seeing what light the supervenience thesis might shed on the physical properties that realise this experience. To that end, if I ask myself what makes me the same person today that I was yesterday, one plausible answer is to relate the continuity of self to self-identity and to link self-identity to memory. If you take away a person's self-identity, their memories, hopes, ambitions, personality, and so on, it seems plausible to say that they are no longer the same person (although they still have a genetic identity that will constrain the person they could become). Nevertheless, they are still in possession of sense of self (patients suffering from total amnesia are nevertheless still selves). In terms of the continuity of the self what matters is that it is the same subjective 'I' that does the remembering and believes itself to have been the perspective from which the remembered events were experienced. Having made memory the key to self-identity and self-identity the key to the continuity of the self, we may then ask how this relates to epiphenomenalism. Employing the supervenience thesis we may answer that those memories supervene on certain neural states in the brain and therefore conclude that it is these neural structures that are the key to the continuity of the self. Although subjectivity may be the essence of the self, subjectivity alone cannot account for the continuity of the self. Since to be subjectively aware is always to be subjectively aware of something, it must be the case that it is the content of our subjective awareness, particularly that which relates to past experiences, that accounts for the continuity of the self. The memories that account for the continuity of the self, however, cannot be viewed as constitutive of the self. Unless they are consciously held, memories are nothing more than (if one subscribes to a connectionist theory of mind) a set of connection weights, or (if one favours a more traditional FP approach) discrete semantically evaluable neural states. Either way they are not the sorts if things that could provide a physical realisation for the self.

I now want to broaden this discussion, and conclude this chapter, by looking at where this concept of the self leaves us and what work we can expect it to do. By focussing on the experiential elements of selfhood and their emergence from the practical order we have secured some safe ground and protected the self from those sociological imperialists who seek to redefine the self in terms of self-identity. As a topic of study, therefore, the universal sense of self is not one that belongs within sociology, though self-identity most surely does. The self should rather be left to psychologists, and perhaps ultimately neurologists, to investigate. Nevertheless, though the self should not be considered as a sociological topic, it is still of vital importance in forming part of the ontological foundations of the social sciences. Without a self that persists through time there is nothing to link the actions that an individual performs over the course of their day-to-day routine, let alone a lifetime.

One of the questions that Campbell argues sociology should address, indeed the question that he views as the most fundamental to sociological theory, is 'what are the conditions necessary for accomplishing individual practical action' (Campbell 1996: 157). One of the conditions necessary is the possession of a continuous sense of self (or rather the subvenient base for the possession of a sense of self). I have to be an *I* before I can engage in any action regardless of whether that action is directed towards other people or the natural world. That is to say, although neonates may reach out to grasp objects and attempt to keep moving objects within their field of vision, such movements do not constitute action until they have a developed sense of self. Without such a sense of self they would not be able to make the subject-object distinction that is necessary for true action. Neither is the 'self' described by the sociological imperialists such as Mead capable of true action. The imperialists' self, as we saw from our discussion of Mead, is constituted by the sum total of others' attitudes towards the individual and by the individual adopting the perspective of the generalised other. Such a self possesses no true subjectivity since even 'introspecting' phenomenal states like pain involves taking on the attitude of others towards oneself. Archer comments (Archer forthcoming) that the Meadean self can never engage in a true inner conversation, and that what looks like an inner conversation is really society talking to society through the medium of the individual. The same is true of action. The Meadean self can never act in any real sense since they lack the subjectivity to do so and, therefore, what looks like individual practical action is really society acting on society.

Chapter 8

Rationality

This thesis is not the result of my endeavours qua rational subject in the traditional sense. It is rather the result of physical processes over which I (as a mere epiphenomenal subject) have no control. I have nevertheless presented my conclusions as objective truth claims about the nature of reality. There is a real tension here between the claims that this thesis is both true and that it is the result of non-rational physical processes. If the thoughts that pop into my mind and the words I type on my computer are all the result of blind physical forces, what reason could there be for our accepting their truth (or falsity)? Rationality, it would seem, requires a knowing subject, one capable of weighing up evidence, making judgements and acting on the basis of good reasons. Without such an autonomous rational subject many people would argue we have no reason to trust our own thought processes, let alone the truth claims made by other epiphenomenal subjects. The most important goal of this chapter, therefore, must be to provide some metaphysical justification for the truth claims made herein. I will begin this justification by arguing, contrary to popular wisdom, that the self is unnecessary for the occurrence of rational thought processes (though it may be logically entailed by the semantics of such phrases as 'the exercise of rationality'). This argument hangs on finding a source of objective knowledge that it not itself dependent upon the minds of conscious subjects. To find this source we may fruitfully apply Popper's theory of objective knowledge in conjunction with his evolutionary epistemology.

With this metaphysical picture in place we will be in a position to consider the operation of practical reason in an everyday context. It is one thing to justify the *possibility* of objective knowledge and rational thought processes without the operation of an autonomous self and quite another to explain how the man in the street approaches the day-to-day problems of existence. The social sciences are particularly vulnerable here since their whole methodology is based, to varying degrees, on the idea of a rational actor. Rational choice theory is obviously the exemplar of this methodology, but even those who tend toward seeing individual choices as the result of social forces need at least a minimally rational actor to translate social forces into actions.

Rationality and the self

We have already encountered the view that a self is a necessary condition for the operation of rationality in our discussion of Searle's theory of system causation. Searle claimed that the existence of an irreducible self is entailed by the logical structure of action sentences. When discussing Searle's work I noted that the logical structure of action sentences is rather flimsy evidence for the existence of the self and suggested that we might be better off questioning the logical structure of action sentences rather than inventing entities, properties, or powers that make them true. At this point I want to consider some other reasons why a self is held by many as a necessary condition for the operation of rationality.

Before we can do so, however, we need to take a look at the concept of rationality itself. Specifically we need to consider what it means for an action to be rational and consider the role played by free will (which itself presupposes an irreducible self) in rational action. Searle provides the groundwork for much of this discussion in his treatment of what he terms the Classical Model of Rationality (see Searle 2001: ch. 1). Although this is a model which both he and I reject (but for different reasons) a brief outline of this model will provide some of the necessary background to the following discussion and is a useful entry point to the debate. Though not intended to be an exhaustive account, Searle identifies six elements of the Classical Model that form the background assumptions underlying most modern treatments of rationality. It is worth noting that these assumptions underlie the rational choice model as well as general theories of rationality. As such the following discussion may be used as a vehicle for questioning the rational choice model.

The first element in this list concerns the role of beliefs and desires in causing rational action. We have spent a considerable amount of time discussing this issue in previous chapters as well as Searle's own view (see ch. 3) so a brief summary will suffice. The Classical Model holds that beliefs and desires function causally in the production of action and that an action may be deemed rational when they function in 'the right way'. That is to say, on the Classical Model beliefs and desires function causally by initiating a casual chain which, barring the intervention of external forces, culminates in action. In this sense beliefs and desires are causally sufficient for action. An action is deemed 'rational', according to the Classical Model, if the set of beliefs and desires which caused the action are logically consistent and in accord with the information available to the person. In chapter 4 I argued that although beliefs and desires (where the referents are functional physical states rather than PEs) may function causally, they certainly do not operate in the crude billiard ball style of causation that is typical of the PFP and the Classical Model being considered here. We can, therefore, agree with Searle that rational actions are not caused by beliefs and desires in the sense that beliefs and desires do not provide causally sufficient conditions for action. This, however, is not much of a concession towards his position since I will argue, contrary to Searle, that being caused by beliefs and desires is not a mark of non-rational or irrational actions (even when beliefs and desires are defined in terms of functional states).

Searle views actions where the antecedent set of beliefs and desires are causally sufficient to for the action to be paradigmatic of irrational and non-rational behaviour. Rational action, in contrast, requires a gap between the antecedent beliefs and desires and action in which an irreducible self can freely choose, on the basis of the available evidence and their beliefs and desires, whether or not to perform an action. The behaviour of a heroin addict, for example, is deemed by Searle to be irrational because their behaviour issues out of causal necessity from an antecedent set of beliefs and desires. Thus if an agent could not have acted otherwise their behaviour is treated by Searle as necessarily irrational/non-rational. What makes this claim problematic is that, as I noted in chapter 4, with the exception of pathological cases and reflex actions, agency is never reported as having been entirely subverted. All actions, including those performed under duress or coercion, are accompanied by *an* experience of agency. Someone, for example, who is forced to sign a confession after even the most extreme torture still experiences the movements of their hand as being under their control, even if the *act* itself (the signing of the confession) is outwith their control. Thus an addict, regardless of the strength of their cravings, never experiences a complete absence of free will. Even assuming that all bodily movement is the result if causally sufficient physical conditions, there is surely some difference in the causal history of acts that are accompanied by an experience of agency and those, like reflexes, that are not. If, as seems right, there is some difference, then the act of giving oneself an intravenous injection is different in kind from reflex or habitual action. Searle claims that rational action requires a self which can freely choose to act on beliefs and desires without being causally determined by those beliefs and desires. What is missing from Searle's account is some justification for the claim that some actions are determined by antecedent beliefs and desires while other 'rational actions' require a self to act on those beliefs and desires. OMRists, like Searle, have to treat a subject's experience of free will as indubitable. If an agent sincerely claims to have performed an action from their own volition neither philosophy, psychology, sociology, nor anyone else arguing from an OMRist perspective, can legitimately question this experience. The reason is that OMRists base their claim that human beings have free will on the PE of agency. Academics that attempt to question the veridicality of this experience are ultimately destroying the ontological foundations of their own argument. Epiphenomenalists, of course, are not constrained by such considerations because the foundations of epiphenomenalism do not rest on PE. Thus, OMRists, if they are to maintain a consistent position, must argue that someone with an intense desire to alleviate their cravings, and who believes that taking heroin would be the best way to do so, must still freely chose to act on the basis of their beliefs and desires so long as they experience their actions as voluntary. Within the tradition of OMR it is perfectly legitimate to treat such cases as examples of weakness of the will (to which we shall turn shortly) but there is no justification for claiming that the actions were causally determined.

A further problematic element in Searle's heroin addict example is that there are clearly a good number of heroin addicts who succeed in breaking the habit. If it really were the case that their heroin use was caused by a set of causally sufficient beliefs and desires then it is something of a mystery how anyone ever manages to change their behaviour. By definition all addicts desire the substance or activity to which they are addicted and most are confronted with the object of their addiction on a daily basis. Most addicts also desire to be free from their addictions. Taking this as a simple set of beliefs and desires we might imagine two individuals equally addicted to substance A, both believing substance A to be the object of their addiction, and both equally desirous to be free from their addiction, yet one person taking the substance and the other refraining. The Classical Model is seemingly unable to explain how the same set of beliefs and desires can cause two different actions. Proponents of the Classical Model are likely to respond by arguing that it was the relative strengths of the belief-desire combinations that caused one person to relent and the other to refrain. This style of response, however, is little more than a just so story. I argued in chapter 4 that the ontology of propositional attitudes adopted by the PFP and by the computational theory of mind goes beyond that which can be inferred from experience. Rather, the existence of propositional attitudes is hypothesised by PFPists to account for what appears to be purposeful behaviour. In proposing that it is the relative strength of the belief-desire combinations that caused one person to relent and the other to refrain, the Classical Model adds another layer of supposition to a theory that is already without empirical or experiential foundations.

The second assumption underlying the Classical Model concerns the role of rules of rationality. In itself this issue is of little concern to us here and can be dealt with quickly; it is worth looking at briefly, however, because it highlights a serious flaw in the rational choice model and because it illustrates an important point relating to the role of semantics and syntax in rational arguments. Searle claims that the Classical Model views thought and behaviour as rational when it is guided by, and is in accordance with, the rules of rationality. Here Searle follows through an example of *modus ponens*[1], which advocates of the Classical Model might appeal to in order to justify inferences of the form:

If it rains tonight, the ground will be wet.

It will rain tonight.

Therefore, the ground will be wet. (Searle 2001: 18)

Another example of the rules that supposedly govern rational thought, this time from rational choice theory, is the transitivity of preference ordering, but we will stick with Searle's example for now. Searle's point, with which we can agree wholeheartedly, is that appealing to *modus ponens* to justify an inference leads inexorably to the Lewis Carroll paradox (an infinite regress where further rules are required in order to justify the application of *modus ponens*). The fatal mistake, according to Searle, is to suppose that a valid argument requires any external justification from *modus ponens* or anything else. Rather, it is the semantic content of an argument that guarantees its validity. The primacy of semantic relations over syntactic relations in the justification of rational arguments presents problems for the version of epiphenomenalism being developed here. If epiphenomenalism is true then the brain must be blind to semantics. The brain can be nothing more than a syntactic engine which, if constructed in the right way and fed with the correct input, generates a valid output. There is a certain similarity here between epiphenomenalism and the

[1] Modus ponens are any inferences of the form: 'If *p* then *q*, and *p*; therefore *q*' as in the rain example which follows.

Classical Model. On both accounts it is the syntax that does the causal work. The Classical Model, however, envisages a simple set of rules (of the sort that one might use to programme a computer) whereas the version of epiphenomenalism being developed here leans more towards a connectionist approach to rational thought. The trouble with both approaches is that although one can imagine a connectionist system, neural network or good old fashioned symbol manipulating AI, being constructed such that it produces valid conclusions in the majority of cases, those conclusions still stand in need of justification and the only possible justification comes from the semantic content of the argument rather than its syntactic relations. One could easily programme a computer with the rules of *modus ponens*, for example, and reliably get the output 'the ground will be wet' when fed with the input 'if it rains tonight the ground will be wet' and 'it will rain tonight.' Such a programme could not, however, be used to justify the inference. Prima facie, if the brain is just a complicated syntactic engine which learns to make inferences based on the same sort of algorithmic processes as the aforementioned programme, then there would be no grounds for trusting human thought processes.

A related point concerning the role of the rules of rationality concerns the putative gap in which human beings have to exercise their free will and choose to make an inference. (This is the same gap that we discussed in chapter 3 between the antecedent causes of an action and its performance.) Searle's point here is that although the premises of an argument may logically entail the conclusion, human beings with free will can still choose not to make the inference. Such a gap, were it to exist could prove problematic for rational choice theory. That is to say, unless beliefs, desires, preferences etc, interact in the crude billiard ball style characteristic of folk psychology and the Classical Model, rules of rationality can not be relied upon as a means of predicting behaviour on the basis of those beliefs, desires and preferences. Rational choice theory makes predictions based on the assumption that agents will follow certain rules of rationality, the aforementioned transitivity of preference orders being one example. If Searle is correct that rational behaviour is not determined by the rules of rationality, but is dependent instead upon semantic content which free agents can choose to ignore, then rational choice theory loses much of its predictive powers. If beliefs and desires do not function causally in the traditional sense and rationality is not a matter of following rules (that is, if connectionism is true), then rational choice theory is based on a flawed ontology. Despite being based on a flawed ontology rational choice theory might still be worth retaining if it enjoyed some predictive success. Since, however, along with much of the rest of sociology, economics and psychology (where it is currently fashionable), rational choice theory has not proved its methodological worth by making successful predictions, it might be time to reassess its usefulness.

In real life decision making rational arguments often provides only the backdrop for the decision making process rather than determining its outcome. A good example of this is provided by Margaret Archer who provides a hypothetical example of a student deliberating over their choice of university. Archer first demonstrates how easily transitivity may be broken by the construction of a system of preference ordering based on assigning numerical values to the applicants criteria of 'league position', 'course content' and 'location' – a method that would surely get the approval of rational choice theorists (see Archer 2000: 69-70). I should note that Archer's example is not in the least contrived, as is so often the case with thought experiments designed to disprove a philosophical position. We need not go into the detail of the example, suffice to say that if people really did attempt to make decisions based on the rational choice model then the transitivity of preference ordering would frequently be broken. Archer then notes that in cases when a person ends up with two or more equally desirable options (according to their model) they are not left in the position of Buridan's ass paralysed by indecision:

> She could behave impeccably [by the standards of rational choice theory] and introduce a tie breaker like 'accommodation'. Yet, why should she, for her main concern is to pick a university, not to establish a transitive preference order. So she is just as likely to pour over the brochures, meet some congenial people at one particular Open Day, or to ask her friends for their opinions. None of these are irrational reactions, but neither are they 'rational choices'. (ibid. 70)

Far from determining the applicant's decision we might imagine that the hypothetical applicant uses the technique of preference ordering as a means of organising her thoughts and helping her to reach a conclusion. Indeed it is just as likely that even if one university had scored highest in each of the three criteria that the applicant would still have decided against the 'rational option'. This does not show any flaw in her original choice of criteria and nor does it suggest that she should conduct the exercise again with a new set. Rather, it shows is that real life decision making is more akin to developing a taste for cabbage over Brussels sprouts than a cold algorithmic process of reasoning. Just as it is not irrational to prefer cabbage to Brussels sprouts it is not irrational to prefer one university to another, even if that preference conflicts with what one 'rationally' decides would be the best choice.

This nicely leads us on to the topic of weakness of the will. Weakness of the will supposedly occurs when a person decides that, all things considered, it would be better to do *x* than *y*, it is within their power to do *x* and yet they do *y*.

Plausibly both the student who decides against the 'rational option' and the heroin addict discussed earlier could both be seen as exhibiting weakness of the will. Weakness of the will has long been a problem for philosophers who view actions as following out of causal necessity from antecedent beliefs and desires, so much so that some proponents of the Classical Model have denied its existence. The traditional solution to the problem has not been to abandon the idea of beliefs and desires being causally sufficient for actions, but to find something inconsistent or incomplete in the antecedent set. Thus Davidson, for example, questions the agent's commitment to performing the action he judged would be best, all things considered. Another popular approach, stemming from Aristotle and Aquinas, is to draw a distinction between rationality and emotion and to argue that reason can be led astray by the passions. Such a dualist picture is now thankfully falling out of favour with the recognition that emotions play an ineliminable role in rational judgements (see e.g. Damasio 1994). However, even if it were the case that reason and emotion were separate faculties (which, incidentally, Searle identifies as another feature of the Classical Model) it seems obvious that reason could anticipate emotional reactions and avoid situations which would, like Ulysses having himself tied to the mast (see Elster 1984), lead to weakness of the will. Such situations are common in everyday life. Any smoker who has tried to give up will, assuming they are committed to the project, avoid situations they know will prove testing. The pub, cups of coffee and meal times are just as dangerous to their project as the Sirens were to Ulysses' and are to be treated accordingly.

Although weakness of the will is generally seen as a philosophical problem it is nevertheless of enormous importance to the social sciences. Many of the issues that are of interest to social scientists, particularly in the field of deviance, can be treated as examples of weakness of the will. Criminal behaviour, drug abuse or truancy, for example, are often likely to be seen by the people who engage in such activities to be against their better interest. Although sociology tends not to talk explicitly about weakness of the will it is implicit in explanations that cite such factors as the lack of deferred gratification or peer pressure. One of the more patronising explanations that sociologists peddled for the underrepresentation of working class children in universities, for example, was that the working class tended not to engage in activities that required deferred gratification. Such an explanation relies on the assumption that working class sixth-formers realised that a university education would, in the long term, improve their employment prospects and life chances but were unwilling to defer financial independence (amongst other things) for the duration of a degree. As such this is a classic case of weakness of the will where one acts against one's better judgement due to immediate temptations. Weakness of the will, of course, is a philosophical problem that is not faced by epiphenomenalism – a non-existent will cannot be weak. Although this does deprive the social sciences of one style of explanation it also saves them from having to deal with the unpredictable nature of fickle minds.

The Classical Model, which as we have seen views rationality as being nothing more than the causal interaction of beliefs and desires according to certain rules of rationality, leaves little room for the operation of a rational self, which many view as essential for justification. Trigg, for example, puts the matter succinctly when he says: 'Consciousness is the necessary pre-condition for rational reflection, giving me the chance to make sense of my experiences.

Without a centre of consciousness '*I*' do not exist, and therefore *I* cannot make any rational judgements' (Trigg 1993: 206). Having denied the existence of a centre of consciousness, an '*I*' that is capable of making the sort of rational judgements that Trigg views as so important, we are forced to claim that the work that was thought to have been done by the self is in fact performed by blind physical (in this case neuronal) forces. This is a view that Trigg finds wholly unsatisfactory. Any realist account, according to Trigg, requires that one is able to transcend the system of which one is a part and adopt a detached position from which to reason and observe. Thus a scientist who talks about the physical system of which they are a part implies that they are able to transcend the causal processes of that system and reason about them. Furthermore Trigg notes, 'I cannot admit that this ability in itself is purely physical without once again adopting a detached position and implying that at least at the next level I can talk *about* physical events, without simply producing another' (Trigg 1993: 222). Even the ability to see a system as a system, which is a pre-requisite for any natural or social scientific endeavour, implies this ability. In the context of social science, for example, Trigg points out that to see a culture as a culture or identify customs, traditions, norms and values, is already to have taken a step back from one's social milieu. Even the most banal sociological observation involves relating such things as customs and traditions to their social context and this, Trigg notes, pre-supposes that one is not bound by that social context. Moreover, if the social context under consideration is not one's own then one must become doubly detached and view traditions in relation to their own context rather than from the perspective of one's own cultural background. For Trigg and others only an irreducible self can attain this position:

> Scientific claims to truth gain their power precisely because they are understood in non-physicalist terms, as saying something *about* the world… Once they are conceived as *only* the end product of a complicated neuronal process, they are facts about individual brains and nothing more. (Trigg 1993: 214)

Strictly speaking Trigg is quite correct here that if theories, conjectures and observations were *only* the end product of neuronal processes then they would be facts about individual brains and nothing more. Scientific claims to truth, however, are never only the end product of neuronal processes. Since the brain is in constant interaction with the external world (via the five senses) scientific claims to truth, on a physicalist account, are the end product of complicated neuronal processes *and* the interaction between those processes and the external world. This would of course be accepted by Trigg. Trigg's real point is that if truth claims are only the end product of physical processes then, regardless of the complexity and scope of those processes, they are facts about physical processes and nothing more. It is this broader claim that I want to counter in the latter half of this chapter. I will argue that far from only being facts about physical processes, theories, conjectures, observations and the rest, can be objective truth claims about the external world. Physical systems such as the human brain are, thanks to a long evolutionary history, learning, and the cultural accumulation of knowledge, able to (almost literally) know themselves. By this I mean that the physical system itself is able to 'know itself' in a manner that does not depend upon any conscious experience that may result as a by-product.

Evolutionary epistemology

The typical physicalist response to the above arguments is to appeal to evolution as an explanation for human rationality. It is doubtful, however, whether evolutionary epistemology can shoulder the entire burden.

> People who do not see holes in front of them are liable to fall into them. So-called 'evolutionary epistemology' is built on this insight… Yet it is unclear that the

> whole edifice of Western science, including the latest research into the nature of subatomic particles, can rest on this need for survival. (Trigg 1993: 44)

Clearly the ability to calculate pi to the *n*th decimal place is something which cannot be linked to genes and the need for survival in the same way as the perceptual abilities that enable us to avoid obstacles. Someday soon it seems likely that we will be able to tell a complete story that shows how our genes guide the development of phenotypic traits such as the optic system which subserve this ability. Such a story will not be forthcoming for our ability to calculate pi or theorise about quantum indeterminacy. Such paradigmatic cases of rational thought are simply too far removed from anything faced by human beings in the environment of evolutionary adaptation (the Pliocene era) to be directly linked to genetics. Not only are such abilities unnecessary for our survival, but as has often been pointed out, they may well hasten our demise by allowing us to construct weapons of mass destruction, destroy the ozone layer, or cause global warming. What we need, therefore, is an evolutionary epistemology that allows us to go beyond crude genetic determinism and provides scope for an explanation of rational thought that embraces learning and the cultural accumulation of knowledge.

The quasi-naturalistic theory of rationality that follows is heavily indebted to and relies on Popper's theory of evolutionary epistemology. Before outlining this approach, therefore, we need to take a close look at Popper's contribution to the debate, especially as it is presented in *Evolution and the Tree of Knowledge* (1972b) and *Epistemology Without a Knowing Subject* (1972a). Popper divides reality into three classes or Worlds, the physical world as described by physics (World 1), the subjective world of conscious experience (World 2, what we have been calling PE), and an objective world of knowledge

(World 3). With one important exception Popper's ontology is compatible with epiphenomenalism. Popper views World 2 as causally efficacious and as providing the bridge between Worlds 1 and 3. I will argue, on the other hand, that Worlds 1 and 3 interact directly, without the mediation of World 2. Before suggesting how this may occur, and in so doing providing a sketch of a quasi-naturalist epistemology,[2] we first need to take a look at World 3.

The most obvious candidates for inclusion in World 3 are the *contents* of books, journals, and libraries. These World 1 texts are the physical instantiations of objective knowledge which exists independently of any knowing subject(s). World 3, despite being largely created by human beings, is by Popper's standards autonomous. Thus, for example, although the sequence of natural numbers is a human creation, once created they display properties (such as prime numbers, odd and even numbers, etc.) that are independent of their human creators. Contrary to the protestations of social constructionists, mathematics is more than an internally coherent language game and we can no more choose to alter its axioms than we can choose to make cheap travel possible by changing the laws of gravity. Another important feature of World 3 is that the objective knowledge it contains need not have been produced by the human mind. A nice example of this is provided by Popper who asks us to imagine a computer programmed to calculate tables of logarithms. Once calculated the tables may be published and distributed but, for whatever reason, left unread. Regardless of whether these logarithms ever see the light of day, however, they remain objective knowledge so long as they have the potential to be understood. Thus even if mankind were to become extinct it remains possible that an extraterrestrial being might uncover

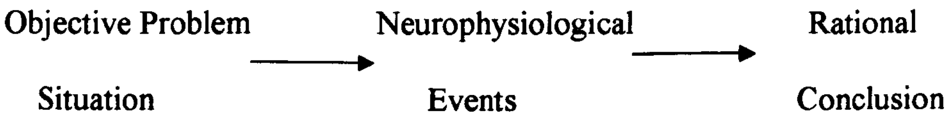
[2] *Quasi*-naturalist because naturalism does not typically talk of World 3 objects.

the logarithms, understand them and put them to use. Objective knowledge need not have been produced by the human mind nor ever be understood by the human mind in order to qualify as knowledge.

The above case also demonstrates the autonomy of World 3 in that the objective knowledge contained in the books of logarithms exists independently of any knowing subject. Nevertheless, since the autonomy of World 3 is such a crucial element in the following argument, I shall present another of Popper's thought experiments (see Popper 1972a: 107-8). This time Popper asks us to contrast two different scenarios. In the first scenario all our machines, tools, and technology is destroyed, along with all our subjective knowledge of how to use those instruments. Nevertheless, in this scenario, libraries remain and so too does our ability to learn from them. The conclusion Popper invites us to share is that after some time, perhaps a few generations, life would return to normal. In the second scenario not only are our machines, tools, technology and our knowledge of how to use those instruments destroyed, but in this case so are libraries, rendering our capacity to learn from books useless. If the second scenario were to transpire we would be pushed back into the Stone Age and it would take as long to recover the lost knowledge as it did to create.

As we shall see below, the autonomy of World 3 can be used to provide the basis for an epistemology consistent with epiphenomenalism. Just as the eye has evolved within a physical environment containing light waves of particular frequencies, so the brain has evolved (both ontogenetically and phylogenetically) within a linguistic environment containing such World 3 objects as states of discussion and states of critical argument. Determinism does not lead to relativism with regard to perception since we can invoke evolutionary arguments to explain the veridicality of perception in a physically determined world. Indeed physical determinism actually makes the veridicality of perception more rather than less understandable since there are no unpredictable or random events to intervene in the causal chain running from object to percept. Similarly, free will has no impact whatsoever on the veridicality of perception. There is simply no way for free will to enter into perception except in the choice of where to look. As has been pointed out before, one cannot raise one's hand to one's face and choose not to perceive it. In what follows I will argue that the same goes for rationality. Until quite recently one of the main arguments for the existence of God, and against Darwinian evolution, was that a structure such as the human eye could not have evolved without divine intervention (an argument that is sadly still being peddled by some creationists). In a similar vein there are several aspects of rationality that seem to require the exercise of a rational agency, judgement, the weighing up of evidence, etc. In what follows I want to draw a parallel between the evolution of the eye, and structures like it, and the development of human rationality. The crucial difference, and this is where Popper's ontology comes into its own, is that while structures like the eye evolved solely in World 1, the environment in which our ability to reason has evolved is comprised of both Worlds 1 and 3. Very crudely then, the operation of rationality reduces to the following causal chain:

Fig. 8.1 A quasi-naturalist account of rationality



Objective Problem Situation → Neurophysiological Events → Rational Conclusion

This quasi-naturalist explanation of rationality is likely to be met with the claim that it is no explanation at all. All that has been done, the objectors will claim, is to take the epistemological and metaphysical problems associated with rationality and to bury them in a box marked 'neurophysiological events'. Now I quite readily admit that there is as yet no convincing empirical evidence to flesh out this account. Nevertheless if we contrast this account with the OMRist account the incompleteness of the quasi-naturalist account is brought into perspective.

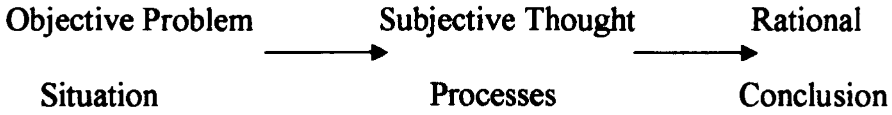


Fig. 8.2 the OMRists theory of rational thought processes

The OMRist faces the same epistemological and metaphysical problems as the quasi-naturalist but instead of a box marked 'neurophysiological events' they insert a box marked 'subjective thought processes' or 'the operation of the self'. Not only is the quasi-naturalist account no worse off with regard to its explanatory power, but it has the advantage over the OMRist story in that future empirical discoveries may one day flesh out its claims. This is decidedly not the case with the OMRists' claims. For reasons that we have already discussed, the irreducible mental states and the irreducible self that the OMRists claim are doing the work will never be amenable to scientific investigation since to do so would require their reduction.[3]

Although we cannot flesh out this claim with empirical material we can help make it seem more intuitively plausible and deal with at least some of the metaphysical problems of justification it generates. One of the issues we touched

[3] See Dennett (1993) for an elaboration of the inadequacy of explanations that cite an irreducible self or irreducible conscious experiences.

upon earlier concerns the supposed inadequacy of evolutionary epistemology in explaining our reasoning abilities. I cited Trigg in connection with this issue who, although happy to admit that the ability to avoid obstacles may have an evolutionary explanation, was dubious about whether evolutionary epistemology could be used to explain the achievements of modern science and in so doing provide a justification for scientific claims to truth. I have already admitted that the specific achievements of modern science are too far removed from the environment of evolutionary adaptation to have a direct evolutionary explanation. Nevertheless, it is important to note that it is not the specific achievements of modern science that require justification (many of which are likely to be well off the mark anyway). Rather, it is the reasoning process that constitutes scientific method. Here the important issues relate to providing a proper grounding for our beliefs. As such we need only show that evolutionary epistemology provides an explanation for why we should be concerned with issues of truth (and falsity) and evidence. This is a much more manageable task for evolutionary epistemology for while our ability to calculate pi is unnecessary for our survival the ability to predict the time of day most conducive to a successful hunt, or when and where we can safely get clean drinking water, most certainly is. Although their subject matters could not be further removed, the same issues of truth and falsity and grounding beliefs in the evidence apply to both hunter gatherers and contemporary scientists. Moreover, as soon as language began to emerge early humans had to interact with World 3 as well as World 1. Thus World 3 became a part of the environment of evolutionary adaptation when humans first communicated thoughts such as 'there is water over there' or 'such and such an animal is easiest to hunt at dawn.' Those who were adept at interacting with World 3 would have gained a huge evolutionary advantage over those who could not. The important point about rationality is that it is not, and indeed cannot, be defined relative to any context. While the eye evolved to register light waves of specific frequencies no such constraints operate on rationality. We could not have evolved to apply standards of rationality to anticipating the escape route that our prey is likely to take and not be able to apply those same methods to constructing iron tools, bridges, or particle accelerators.

As human rationality evolved in conjunction with our linguistic abilities we not only confronted a pre-existing World 3 but we began to actively create our own World 3 environment. A good example of this self-created environment comes from Popper who describes a meeting with Bertrand Russell:

> Many years ago I visited Bertrand Russell in his rooms at Trinity College and he showed me a manuscript of his in which there was not a single correction for many pages. With the help of his pen, he had instructed the paper. This is very different indeed from what I do. My own manuscripts are full of corrections – so full that it is easy to see that I am working by something like trial and error; by more or less random fluctuations from which I select what appears to me fitting. We may pose the question whether Russell did not do something similar, though only in his mind, and perhaps not even consciously, and at any rate very rapidly. For indeed, what seems to be instruction is frequently based upon a round about mechanism of selection…
>
> I suggest that we might try out the conjecture that something like this happens in many cases. We may indeed conjecture that Bertrand Russell produced almost as many trial formulations as I do, but that his mind worked more quickly than mine in trying them out and rejecting the non-fitting verbal candidates. (Popper 1987: 145-6)

We have here a nice example of the way that an individual mind creates for itself an environment of conjectures and objective knowledge within which it can reason. Though Popper claims to be producing almost random fluctuations this can not be quite right. The story of the thousand monkeys at the thousand typewriters comes to mind here, though one can imagine a situation where the brain produced random combinations of words and concepts for the scrutiny of the self, such a method of thought could not possibly generate the vast quantity of more or less coherent thoughts we all experience. Rather, it must be the case that fed with the appropriate material the mind/brain generates a series of novel re-combinations all of which are to a certain degree logically coherent. It may well be the case that most of the trial combinations are edited or discarded well before they become conscious. Those that make it to the level of consciousness, and especially for those that survive to be recorded, become World 3 objects and thereafter are themselves the subject of appraisal and re-combination. I should note that it is not in virtue of their being conscious that such recombinations become World 3 objects. We know from empirical research on neuronal adequacy (neuronal adequacy relates to minimum time span a neuronal state must persist if it is to generate conscious awareness) that those states which generate conscious awareness involve proportionately larger neuronal assemblies and remain active for a proportionately longer period of time than those that do not. This, it seems plausible to suggest, makes it more probable that they will become the objects of other neuronal states 'attention'.

Whether innate or learnt it seems likely that the brain is wired such that it produces grammatically correct sentences without the intervention of a conscious self. I suggest that something similar goes on in the brains of those who seek to address such questions as the problem of induction, of structure and agency or of quantum indeterminacy. Their prior knowledge of and experience within their respective disciplines wires their brains such that only relatively plausible solutions are generated. The more prior knowledge and experience one has of a discipline the more likely one's initial conjectures are to hit the mark. Anyone who has considered an issue for a prolonged period of time will recognise this from reviewing early drafts of their work which, in retrospect, appear laughable to their author.

The phenomenology of rational thought lends some support to this hypothesis. When we approach a problem we begin by feeding the brain with a good number of theories, conjectures, evidence, etc. from which, without conscious intervention, thoughts and ideas that are an amalgam of and a reworking of previous ideas pop out. We have no phenomenal experience of the process of reasoning only of its consequences. The self passively experiences the scales of rational judgement which have been weighted by some non-conscious process. What does one do when one engages in rational decision making? Speaking for myself I do not *do* anything. I experience certain thoughts, some negative others positive, some trivial or irrelevant, and a good many that are emotionally charged, and after some time I experience a decision or a conclusion popping into my mind without my having *done* anything. If there really is a self weighing up the alternatives then it is a self of which I am unaware.

Perhaps one day connectionism, or some similar discipline, will elucidate the ways in which non-rational bits of matter are able to make novel scientific discoveries, create works of art and develop theories of evolutionary epistemology. Such an achievement, it seems, is some way off, and the mind boggling complexity of any imaginable solution beguiles the imagination into conceiving of the project as an ill-conceived flight of fancy. It is much simpler and more intuitively plausible to conceive of reason as a faculty belonging to and exercised by an autonomous self. By doing so, however, we surrender any attempt at explanation by invoking some mysterious 'mind-stuff', to borrow Dennett's phrase. As this is a work in philosophy and social theory we need not concern ourselves with the empirical details of a theory of evolutionary epistemology, nevertheless, if possible it is incumbent upon us to present our theories in a manner which is both intuitively plausible and logically coherent (the two not necessarily being close bedfellows). The following quote from Campbell goes some way to achieving this task.

> A major empirical achievement of the sociology of science is the evidence of the ubiquity of simultaneous invention. If many scientists are trying variations on the same corpus of current scientific knowledge, and if their trials are being edited by the same stable external reality, then the selected variants are apt to be similar, the same discovery encountered independently by numerous workers. This process is no more mysterious than that all of a set of blind rats, each starting with quite different patterns of initial responses, learn the same maze patterns, under the maze's common editorship of the varied response repertoires. (Campbell 1987: 71)

Where scientists (and rats) have an external physical reality to edit their conjectures, social scientists, philosophers, mathematicians, logicians and (to an even greater degree) artists, are often several floors removed from the editor's office. Though ultimately all objective knowledge must be grounded in external reality, the above disciplines frequently have to deal exclusively with World 3 objects embodied in World 1 texts. In such cases, despite their being far removed from physical reality, they are able to test their theories and conjectures against the autonomous world of objective knowledge (World 3). There is a danger common to all disciplines that operate solely with World 3 objects embodied in World 1 texts of becoming little more than internally coherent language games with little or no connection to the physical or social realities they purport to be about. It is important to recognise that World 3 may be autonomous without being true. A good example of this comes from science fiction writing. A science fiction writer is constrained by the pre-existing cultural, technical and physical 'laws' of their genre and these laws in part determine the realms of possibility. Previous films, series, and books become the autonomous World 3 environment which 'edits' the author's storylines. The danger is that academics have the potential to become constrained by their own ‘fictional’ World 3 environment in just the same way as science fiction writers are constrained by theirs. Once texts, authors, or theories, gain currency within a discipline and are accepted as true, they exert a power on future writers that is analogous to that of the physical world on those working in the natural sciences. There are several schools of thought within the social sciences that might be thought to have fallen into this trap. Certain strands of critical and deconstructive social thought, for example, might be seen as little more than internally coherent language games. One candidate for this line of criticism is Marcuse’s *Eros and Civilization* (1998). Marcuse’s seminal work is based almost exclusively on Freudian and Marxian theories. Marcuse’s aim was to unify Freudian and Marxian theory in order to develop a critical theory of contemporary society. How successful Marcuse was in this endeavour is not an issue that I wish to tackle here, the point at issue is that Marcuse was operating almost exclusively within a World 3 environment (there is no rigorous empirical work contained in the text). There is a fact of the matter about whether these two theories can be unified, but it is an objective fact that is independent of the truth of either theory. The works of Marxists and Freudians (World 3 objects) are in this case the editors of Marcuse’s theories and conjectures and have far more authority than an independent physical or social reality. Given the well known speculative nature of Freud’s work, his lack of rigour in documenting case studies, and his reluctance to consider the possibility that his patients might be suffering from organic illnesses (see Webster 1995), one of the foundations for Marcuse’s project is of dubious worth. Couple this with the dogmatism of the Marxism of the period and Marcuse’s World 3 environment does not look conducive to the discovery of truth. Now, of course, *Eros and Civilization* is a World 3 object and forms part of the environment in which contemporary critical theorists reason.

The existence of an autonomous World 3, in conjunction with an evolutionary epistemology that explains how we can come to have knowledge of the content of World 3, provides the groundwork for a realist perspective on rationality. This then allows us to counter the relativism/social constructionism of certain currents within the sociology of knowledge such as the 'strong programme', which views all knowledge as normative. Academics and students within the social sciences (and sociology in particular) have become obsessed with issues of self-reflexivity and interpretation that have emanated from within postmodernism. No longer do they see themselves as dispassionate observers who can at least attempt to understand and report about other cultures (even those that are similar to their own). Such an attitude begs the question of why bother continuing to work within the social sciences. After all, if one abandons the belief that the social world is open to objective investigation one may as well take to novel writing.

Epiphenomenalism does lead to scepticism about the correspondence between PE and objective reality, since if epiphenomenalism is true it would make no causal difference if there were no correspondence between the content of PE and objective reality. However, epiphenomenalism is consistent with realism and is straightforwardly compatible with evolutionary epistemology. The only caveat is that the realm of objective knowledge is restricted by epiphenomenalism to that which is contained in Worlds 1 and 3. World 2 may also contain knowledge about objective reality, but since such knowledge is, according to epiphenomenalism, inaccessible and cannot be tested against objective reality, we may never be in a position to judge how closely the content of PE mirrors objective reality. The supervenience thesis does lend some support to the hypothesis that the content of PE covaries in a systematic and reliable way with external reality, but it cannot guarantee correspondence. To explicate, evolutionary epistemology guarantees the veridicality of perception for a person with normal vision and sufficient light (at least in the majority of cases). That is to say, we can be sure that if a person directs their gaze towards a tree their brains will register certain of the objective features of the tree. Knowledge of such objective features then enables the person to, for example, walk around it without bumping into it. The supervenience of PE on neuronal events also guarantees that every instantiation of a given neural event will invariably be accompanied by the same PE. Thus we can be sure that every time a person perceives a tree their brain will instantiate a neural state with tree related content. We can also be sure that every time a person's brain instantiates a neural state with tree related content they will experience a similar PE. What we can not be sure about is that their PE corresponds in any way to the objective features of the tree they are perceiving. The conclusion then is a restrained scepticism about PE, but confidence about the possibility of objective knowledge. Epiphenomenalism, therefore, lends no support to relativism as long as that knowledge has been tested (in an appropriate way) against Worlds 1 and/or 3. I will leave the worrying about what the 'appropriate way' is to philosophers of science. The greatest threat to rationality, therefore, comes not from epiphenomenalism or the denial of the existence of an autonomous self, but from those who embrace the contradiction and attempt to reason away reason.

Chapter 9

Action, Social Action, and Epiphenomenalism

In previous chapters we have sought to explain what man is and have set aside questions about what man does. In this final chapter I want to explore the latter question and consider how we should approach the study of action given the conclusions drawn earlier. Action, it seems fair to say, is the most important concept within the social sciences. Without a sound understanding of action, including a basic definition of the concept, an account of the conditions necessary for action, and an appreciation of how action relates to interaction and social action, all conduct would be unintelligible. That said, however, it seems unlikely that the traditional concept of action, being (loosely) conduct which is both voluntary and subjectively meaningful, can survive within any discipline that embraces epiphenomenalism. Once subjective states are conceived as causally inefficacious, and the distinction between voluntary and involuntary conduct is rendered spurious, the distinction between action and behaviour (the latter being defined, again loosely for now, as reactive responses) collapses.

Given the importance of the concept of action, therefore, our primary goal in this chapter has to be to reconstruct the concept in a manner which is compatible with an epiphenomenalist ontology. Equally important, however, is to defend this reconstructed concept against those who doubt the usefulness of the concept or even the existence of action as a subject matter distinct from social action. Here I refer to those that have bought into what Campbell calls the paradigm of social situationalism and the myth of social action. Social action, defined as action which possesses a social meaning, has, according to Campbell, come to displace the more general concept of action (which, as I stated above, is defined as conduct which is voluntary and subjectively meaningful). The upshot of this shift in perspective is that sociologists no longer have the tools at their disposal to provide a causal account of how and why people behave the way that they do. Since the ultimate goal of any science is causal explanation, this is a worrying trend that I shall attempt to counter in this chapter. Doing so will involve, somewhat paradoxically given the subject matter of this thesis, making mind matter more in the ontology and methodology of the social sciences. Specifically it will involve making actors' accounts that are traditionally thought of as referring to PEs central to the study of action. That is to say, I will argue that accounts which cite such things as beliefs, hopes, fears, desires, interests, motivations and so on, should be the primary (though by no means exclusive) resource for the study of action. I have previously hinted that the referent of such mentalistic terms is not, as has previously been assumed, PEs. Rather, if epiphenomenalism is true, the referent of mentalistic terms is the content of causally efficacious neural states. Thus the stated aim of 'making mind matter more' is perhaps a little misleading since what I mean is that the language traditionally associated with the mind should be made to matter more.

Before we get ahead of ourselves let's begin by outlining the ontological implications of epiphenomenalism for the traditional concept of action. At the ontological level, since this thesis is concerned only with the individual and does not consider the interaction between man and society, we will confine our discussion to the proximate causes of action. That is to say, we will consider what internal causes are implicated in causing actions and ignore the wider question of how these internal causes came to exist in the individual at any given time. Thus we will consider what internal states *y* of an individual caused action *x*, and not what social or biological forces were causally effective in bringing about state *y*. The most obvious implication of epiphenomenalism for action is that all actions are caused by antecedent and synchronous physical states and not by the PEs that accompany them. PE, being entirely inefficacious, is thus irrelevant for a causal study of action. As I previously noted, however, despite PEs themselves being irrelevant to the causal study of action, the language that is thought to describe those experiences is essential.

This obvious conclusion presents us with our first quandary. Traditionally action has been differentiated from behaviour by reference to their respective causes. Action is typically defined as voluntary or purposeful conduct in contrast to behaviour which is viewed as reactive or (sometimes) habitual conduct. Despite the long standing debates over free will versus determinism and reasons and causes that have made reaching a precise definition of voluntary action notoriously difficult, it is nevertheless the case that there is a common consensus that voluntary action is caused by some conscious act of 'will'. Since epiphenomenalism offers such a radical alternative to this debate we need not revisit this old battleground and consider what is meant by 'willed' conduct. Suffice to say that in all the traditional definitions of voluntary action PEs of a particular kind always play a causal role. That is to say, whether one is a determinist who views actions as just one more link in a relentless causal chain, or one believes in the existence of an irreducible subject who causes actions by performing an act of will, action is always differentiated from behaviour by having the right kind of causal history and phenomenal states of a particular kind always form part of this causal history. If epiphenomenalism is true, however, then it would seem to render this action-behaviour distinction meaningless. After all, if all movement is caused by internal physical states then what difference does it make if some of those states are accompanied by epiphenomenal experiences?

Despite this I will argue that the action-behaviour distinction is worth retaining. Epiphenomenalism notwithstanding, there surely is a difference in kind between a reflex knee jerk and the 'deliberate' act of kicking someone in the shins. The challenge is to find a means of differentiating between the two classes in a way that does not rely on giving PE a causal role. One means of doing so might be by retaining the classificatory emphasis on the causal history of the act, but to emphasise the causal role of antecedent physical states in place of the traditional focus on phenomenal states. So, for example, we might say that movement is action if it is caused by physical states that are themselves the subvenient base for PEs such as intending, trying, concentrating, etc. This then makes PE causally relevant in predicting and explaining actions. Things, however, are not quite that simple. Although I have previously claimed that there is a systematic and reliable relationship between the occurrence of certain phenomenal states (such as pain) and certain behaviour patterns (like withdrawal), this relationship can not be exploited for the purposes of either defining the difference between action and behaviour or for prediction and explanation. The reason for this is quite simply that because PEs are epiphenomenal their occurrence cannot be used to demonstrate the existence of a causally effective subvenient base since to do so they would require some causal powers (indicating the presence of causally efficacious properties is itself a casually efficacious property). The claim, therefore, that PE is causally relevant must be read as shorthand for the following: PE covaries in a systematic and reliable way with neural states that are causally efficacious in bringing about actions. Thus pain covaries with withdrawal because pain covaries with a neural state/mechanism that causes withdrawal. Similarly, both pain and the subvenient base for pain covary in a systematic and reliable way with certain behavioural manifestations of pain such as wincing, crying, screaming and expressions such as 'ouch' and '(insert your favourite expletive) that hurt'. Thus we can use both a person's observed behaviour and their accounts as the raw material upon which to make the action-behaviour distinction. Throughout the rest of this chapter any use of language that implies the causal efficacy of PE should be read as shorthand for the above.

This classification of behaviour and action has certain behaviourist overtones that some will find objectionable. Objectionable it may be but some form of behaviourism seems to be the inescapable consequence of epiphenomenalism. Though I can do little more than flag up the issue here it also seems that epiphenomenalism lends considerable support to a number of behaviourist doctrines. Perhaps the most striking example concerns the behaviourist theory of how people learn the use of words referring to sensations and mental states. Wittgenstein has done much to dispel the idea that people learn the meaning of terms for sensations on the model of 'object and designation'. In his famous beetle in a box analogy Wittgenstein showed the absurdity of the idea that people know the meaning of words like 'pain' only from their own case – by naming a private object or picture (see Wittgenstein 1968). Wittgenstein's argument, which we shall encounter again below, is that such private pictures have no relevance in the grammar of the expression of sensations since there is no way of ensuring that the same mental picture is present in the minds of different users of the language or, indeed, that the picture remains constant in the mind of a single language user. His conclusion was that the meaning of mental states (such as hoping, wanting, believing) and the terms for sensations cannot be learned exclusively from introspecting one's mental states and naming what one finds. Instead, part of the meaning of these terms must be derived from, and learned with reference to, their public use. If epiphenomenalism is true then it turns out that no mental 'pictures', including such things as the PE of pain, hoping, wanting, believing, etc., have a causal role in our learning the meaning and use of mentalistic terms. Furthermore, once learned we cannot even use these public terms as a means of giving expression to our subjective states. To do so, of course, would again be to ascribe a causal role to PE. Thus it would be impossible for a person to learn the meaning of the word 'pain' by observing other people's pain-behaviour and then to use that knowledge to label their own experience of pain. Epiphenomenalism then would seem to lend considerable support to behaviourism generally and the doctrine of logical behaviourism in particular. The central tenet of logical behaviourism is that the meaning of mental terms is exhausted by the observations that are used in the determination of their use. That is to say, once one has learned the meaning of a word by observation of the external world and the behaviour of others, there is no 'surplus meaning', PE or Wittgensteinian 'picture' which contributes further to its meaning. Though Wittgenstein's arguments, and those of the logical behaviourists, were designed to deal predominantly with the meaning of mental terms, if epiphenomenalism is true then it extends to the use of all words including, for example, nouns and proper nouns.

The behaviour-action distinction can still be usefully retained even in the absence of a voluntary-involuntary distinction. It seems fair to suppose that behaviour and actions are the culmination of two quite different kinds of causal history. Though we may be in the dark about the specific details of these causal histories (and will likely have to wait some time for the special sciences to fill the gaps in our knowledge) knowing from which kind of causal history a person's conduct results, and labelling that conduct action or behaviour accordingly, will remain an essential tool in predicting and explaining that conduct. The most effective means of illustrating this point is with an example. Consider a medical researcher who is interested in why people engage in conduct to which they are addicted or to which there is a possibility that they might become addicted. In the first case, of conduct which is caused by addiction and hence may be termed behaviour rather than action, there will be a physical cause that is different in kind from conduct which only has the potential to become addictive and hence constitutes true action. Being able to tell the difference between action and behaviour, so defined, will be essential if the medical researcher's goal is to develop ways of preventing people from engaging in activities that have the potential to become addictive (such as the first cigarette, shot of heroin or visit to a casino). In the case of conduct caused by addiction, there may often be behavioural cues that allow researchers to identify the cause of the conduct in question (such as the visual manifestations of withdrawal symptoms). Often, however, it will only be by listening to an actor's account of their conduct that the researcher will be able to gauge the level of addiction and hence identify the cause of the actor's conduct.

Social situationalism: the paradigm

Having briefly considered some of the more obvious ontological consequences of epiphenomenalism for action theory I now want to consider the contemporary approach to action theory within sociology. Here we will focus much of our attention on Campbell's recent (1996) work *The Myth of Social Action* in which Campbell examines what he calls the dogma of social situationalism. Campbell describes his work as a critique of a critique. It is a critique of social situationalism which is itself a critique of traditional Weberian action theory. Campbell's aim is to defend Weberian action theory against the contemporary obsession with interpersonal communication, and the interpersonal construction of meaning, over intrapersonal processes in the understanding of action. That is to say, in all areas relating to the ontology and methodology of action theory, social situationalists place greater emphasis on the social dimension of action than on its subjective and private dimension. Although an epiphenomenalist ontology does not sit easily with Weber's (since Weber was certainly an OMRist) it is nevertheless the case, or so I shall argue, that Weber was at least looking in the right place for an explanation of action. Campbell claims that '...the key feature of classical action theory was the attempt to explain the conduct of individuals via an understanding of the meanings which their actions have for them' (ibid. 30). Now although epiphenomenalists' approach to meaning is radically different from that of Weber or the classical tradition generally, they would still argue that classical action theory is a fundamentally sound methodology. That is to say, if the version of epiphenomenalism set out in the previous chapters is correct, to understand action one has to look primarily to the internal states that are the proximate cause of action. Thus, classical action theory is fundamentally correct to seek an explanation for action that refers to the internal states of the actor. Where an epiphenomenalist approach and that of the classical tradition diverge is over the issue of which internal states are causally efficacious. Where the classical tradition views PE as the irreducible cause of behaviour and action, epiphenomenalists view those states as causally inert. Nevertheless, as I argued in the introduction to part II, if we treat actors' accounts as referring to causally potent neural states rather than causally inefficacious phenomenal states, the language of folk psychology can still be usefully employed as a means of predicting and explaining behaviour. It is because actors' accounts are causally relevant in this sense that the methodology of classical action theory is fundamentally sound. As such Weber's methodology is one that can be moulded to suit an epiphenomenalist ontology in a way that the various methodologies comprising social situationalism cannot.

Because social situationalism privileges the interpersonal and intersubjective it fails, according to Campbell (see ibid. 153), to address the question of why action happens at all. That is to say, social situationalism fails to explain why a person chooses action over inaction or one course of action out of the myriad of possibilities open to them. I will, of course, expand on this point later, but the important distinction is that classical action theory correctly identifies the cause of action as being 'in the head' of the actor and, therefore, on this issue at least, is not antithetical to my own position. Social situationalism, in contrast, is typically not interested in what is going on in the head of the actor and consequently lapses into observer relative *descriptions* of social interaction. It is important to stress that this is not a mere methodological quirk of social situationalism: it is a problem that goes to the ontological heart of the paradigm and nothing less than its replacement will get sociology out of its current predicament.

Campbell is on the whole very charitable to social situationalists, declaring on page one of the critique that sociology should be a broad church 'which permits its members to study any aspect of social phenomena in any manner they wish.' This is, perhaps, too charitable an approach. I have for a long time wondered quite what the point of sociology is, or rather what the point of contemporary sociology is. Historically sociology was a discipline that sought to *causally* explain the structure of human societies as well as the interaction of its members and groups. Until recently the ontological foundation for this project was a concept of man as a creative and active subject who, by their actions, helped to shape the social world of which they were a part. That is not to say that the early sociologists did not place great importance on social and economic forces, for clearly they did, but these forces, where active, were always seen as operating through the medium of active subjects. Though many of those explanations now strike us as rather simplistic, naïve and in many cases ideologically prescribed, the endeavour was both laudable and defensible. The belief that one can understand (as well as predict and explain) actions from the inside made these important projects which, if successful, could have a real impact on peoples lives. Contrast the work of the early pioneers with the fruits of conversation analysis, ethnomethodology, dramaturgical analysis, all conducted by practitioners that are so consumed with postmodernist inspired anxieties that little empirical work gets done and one wonders if one is looking at the same discipline. Now in one sense Campbell is quite correct, we have no right to tell people how to do sociology, but unless the ultimate goal is a causal explanation or in some way helps to enhance the lives of the subjects of sociological research (and ultimately fund that research) then sociology will find it increasingly difficult to justify both its existence and its funding.

As we shall see, the defining feature of the dominant paradigm within sociology (social situationalism) is the privileging of interpersonal processes over intrapersonal processes in the understanding of action. This has the methodological consequence of making interaction the focus of empirical research and social theory and fosters a disinterest towards actors' points of view. Prima facie this development within sociology would appear to be well suited to an epiphenomenalist position. After all, if mental states are causally inefficacious why bother with the old style Weberian interpretative method, far better to focus on interpersonal communication which we know to be causally potent and is relatively untouched by epiphenomenalism. Interpersonal communication is, of course, untouched by epiphenomenalism because, despite often making references to PEs and propositional attitudes, the medium of communication (sight, hearing, etc.) is not dependent on PE for its efficacy. That is to say, the causal mechanisms implicated in one individual asking another to 'pass the condiments', and the other doing so, (an example that Bhaskar (1998: 105) erroneously believed falsified epiphenomenalism and physicalist determinism) are explainable in microphysiological terms which make no reference to PE. One might expect therefore, that the methodology of social situationalism would be ideally suited to an epiphenomenalist ontology since it makes little if any use of phenomenal states. As I have previously noted, however, although PEs themselves should have no place in causal explanations, mentalistic terms and the language of folk psychology certainly do. In chapter 4 I argued that folk psychology remains a useful predictive tool so long as we interpret the referent of folk psychological terms to be causally efficacious neural states and not their epiphenomenal consequences. Indeed, barring some great and improbable leap forward in neuroscience, folk psychology and the interpretative method will remain the *only* means of gaining access to causally potent neural states. Thus the methodology of classical action theory is entirely consistent with an epiphenomenalist ontology such as my own.

Before beginning our discussion of contemporary action theory within sociology – or rather the lack of it – it might be worth reminding ourselves of the classical position on action. Here I will confine myself to a brief discussion of Weber's original position and the distinction he drew between action and social action, before outlining Campbell's claim that this has been usurped by social situationalists. For a detailed discussion of Weber and the classical tradition, as well as the influences that led to its revision, I would encourage the reader to refer to Campbell's text. Weber, in common with most philosophical action theorists, draws a distinction between behaviour, viewed as involuntary or reactive responses (often including habitual behaviour), and action, which is differentiated from behaviour by being both voluntary and subjectively meaningful conduct. Social action, in Weberian terms, is then defined as action which is orientated towards others. For Weber the defining characteristic of all action was that the act possessed a subjective meaning for its executor. Social action is, therefore, no less subjectively meaningful than action and should properly be conceived as a sub-class of the more general category of action.

Campbell's contention is that action, so defined, has all but disappeared from both social theory and the empirical practice of the discipline and has been replaced by the concept of social action. Despite this transition having frequently been made (erroneously) in Weber's name, the modern concept of social action, according to Campbell, bears little relation to Weber's original definition. For social situationalists the defining characteristic of social action is not, as it was for Weber, action that is orientated towards others, but instead is action which possesses a social meaning. Having made this claim, which we shall examine in detail below, situationalists extended their argument to claim that, since according to their ontology all action possesses a social meaning, all action is necessarily social action. In this way action was dropped from sociology's ontology and with it went the interpretative method and any interest in subjective states.

Although the historical details of how, when and by whom this transition was effected are of little interest to us here, the arguments proffered by the situationalists to support their contention that all action is necessarily social are worthy of consideration. The first argument to be considered here is the claim made by situationalists that mental states, as they have been traditionally understood, either do not exist, or if they do are inaccessible. The basic tactic employed by situationalists is to relocate meaning in social situations, 'situationalists deny that meaning resides in the minds of actors, insisting instead that it is located in the social situations and it is shared, "intersubjective" and contextually determined and manifest' (ibid. 43). As Campbell correctly notes (see ibid. 45, 53), strictly speaking meaning can only be located in the minds of individuals (though I would argue that such things as texts or theorems remain *meaningful* in the absence of any conscious observer). This ontological truism highlights something of a contradiction within the situationalists' paradigm, for it is logically impossible for meaning to be located only in social situations and for it to be shared and intersubjective. For meaning to be shared and intersubjective the same meaning must be present in the minds of at least two individuals and, therefore, the claim that meaning resides in social situations rather than the minds of individuals must be false. Consequently the claim that meaning resides in social situations must 'be taken as shorthand for statements which specify the precise circumstances under which given meanings are registered in the minds of particular individuals' (ibid. 45). The longhand version of this claim might be that in denying that the mind is the location of meaning situationalists mean to suggest that the origin of meaning comes from social situations rather than being generated in the mind. If this is what the situationalists have in mind, however, then it is easily falsified by the existence of misunderstandings. If meaning really were generated by the situation and then somehow 'transmitted' into the minds of those present, or read from the situation, then one would expect those present to share the same understanding of the situation. The moment that one admits the existence of misunderstandings one has admitted that actors have somehow contributed to or distorted the meaning which the situation has imparted to them. In so doing they have created meaning independently of the social situation and the situationalists' claim must, again, be false. In addition to misunderstandings another obvious fact that can not be dealt with from within the situationalists' paradigm is the existence of original thought. This latter problem will be dealt with below.

One of the arguments that situationalists use to justify their ignoring subjective mental states is what Campbell calls the argument by denial. This amounts to the claim that sociology can legitimately ignore subjective states and intrapersonal processes because these states either do not exist or are accessible *only* through first person knowledge and hence are inaccessible to the researcher. One of the main sources of support for this thesis was drawn from Wittgenstein and the post-Wittgensteinian philosophy of Winch and others. In particular it was Wittgenstein's rejection of the idea that the meaning of a word is derived from the thing, or mental 'picture', that a word stands for that situationalists latched onto. However Wittgenstein's argument that the meaning of mentalistic terms does not derive from the mental 'pictures' to which words seem to refer, but instead derives from their use, has no implications for the debate surrounding the causal role of subjective states (at a methodological or an ontological level). The fact, for example, that the verb 'to imagine' could not exist in an absolutely private language[1] has nothing whatsoever to do with whether or not (for example) a climber's act of imagining themselves successfully completing the crux of a climb is causally efficacious in helping them to do so. Indeed we can take this argument a step further, for while it is certainly the case that in order for a climber to be able to say to himself or another, 'I am going to imagine myself successfully completing the crux of that route', they must make use of language (which is social in origin) the ability to perform the act itself is not dependent on *any* linguistic ability. Moreover, since the ability to imagine evolved in humans well before language, it must be the case that innumerable acts of a similar type were performed well before man uttered his first word.

Since we have already discussed similar issues in previous chapters I do not want to overlabour the point here, but I should like to reiterate the claim

[1] In an absolutely private language the referent of a word would be a mental picture of the object or concept to which the word refers. Wittgenstein successfully demonstrated that such a language

made earlier that although the origin of language is social its use can be entirely private. Thus although social situationalists are correct in their assertion that the genesis of meaning is dependent on intersubjective agreement relating to a situational context, once humans have acquired language they have acquired the ability to transcend its use in the social situation in which it was acquired. If this were not the case then there could never be any original thought since, by definition, original thought involves the independent creation of new meaning. What's more, as we noted in our discussion of Popper, original thought remains meaningful (in the sense, for example, of its being true or false) regardless of whether another conscious subject ever becomes aware of the content of that thought. This point is important because it denies situationalists the escape route (and one that might have particular appeal to social constructionists) of being able to claim that although individuals are able to formulate new sentences or write down new theorems, that these sentences and theorems only become meaningful once they have gained the assent of, for example, the scientific community.

Subjective meaning in action and accounts

In this section I want to contrast what I think are two different classes or types of action. Examples of type I actions are as follows: spinning a prayer wheel, hanging prayer flags, throwing salt over your left shoulder, waving goodbye. These are actions which are unquestionably subjectively meaningful to the actor performing them and, one would assume, where the subjective meaning is causally relevant in an explanation of the actions. They are also the types of

would not be possible because, in essence, there is no means of ensuring the constancy of the mental referent.

action for which a relatively coherent explanation could be provided by the actor if their conduct were questioned. Contrast these examples, however, with type II actions such as chopping wood, cooking dinner, having a nap, taking the car in for a service and so on, and it is not at all clear what the subjective 'meaning' attached to these acts is. They are certainly examples of purposeful behaviour but they do not possess the same sort of readily identifiable subjective meanings as those of type I. It is worth asking why, if someone were questioned about an action of type I, they would be able to respond in a clear and coherent way, but if asked to explain an action of type II, they would likely respond with a look of puzzlement. One reason might be that that the meaning of the second class of actions is embedded in the act itself. Chopping wood is just chopping wood and the meaning of the action is likely to be obvious from the context (i.e. a workplace or a domestic setting). Moreover, the purposes underlying such actions are likely to be universally understood and readily apparent. The subjective meaning that is attached to examples of actions that comprise the first class, in contrast, is in no sense inherent in the act. Spinning what one knows to be a prayer wheel, for example, is a different act from spinning what one takes to be just a piece of wood with marks inscribed on its surface. It is for this reason that one can confidently say that the subjective meaning that the actor ascribes to the act is causally relevant in any explanation of that act. The example also serves to highlight the fact, commented upon by Campbell, that an observer can not assume knowledge of the subjective states of the actor, or what the social meaning of the act might be, without directly enquiring as to the content of the person's mental states. From the perspective of an observer, for example, there is no way of telling whether an individual is spinning a prayer wheel as an act of worship or whether they are a naïve tourist spinning what they take to be just a piece of wood with marks inscribed on its surface. Now it may turn out that the proportion of cases where misinterpretation occurs is insignificant and that researchers may continue to read off the subjective and/or social meaning of an act from observation alone. However, this is an empirical question which is yet to be settled and as such it remains a dangerous assumption (see Campbell 1996: 132-135). The moral is that acts tend only to have a true subjective meaning rather than being just purposeful and intentional when the meaning or purpose is not embedded in the act. There are of course exceptions to this general rule, chopping wood might be an excuse to avoid getting down to writing the next chapter and in this case the meaning would no longer be embedded in the act. Nevertheless I think that this is a useful distinction and exceptions only serve to highlight the fact, recognised by all those working within the Weberian tradition, that the only way of finding out the subjective meaning of an act is to ask the actor.

A second important distinction between the two classes of action is that the meaning of type I has to be learned (through socialisation) in a way that the purpose of type II actions do not. This should not be taken to mean that the ability to perform the second class of actions is somehow innate, or that their execution does not constitute a skilled performance on the behalf of the actor involved. Rather, it means that the type of knowledge invoked in actions of each type is different in kind. Actions of type I require that the individual learns and internalises a shared social meaning, whereas actions of type II require only the internalisation of practical and technical knowledge. This may provide a partial explanation for why sociologists have moved from Weber's wide definition of action towards the modern narrow interpretation that action is behaviour that possesses a social meaning. The reason for this is that many of those actions that possess a true subjective meaning (as in type I), rather than being behaviour where the purpose is embedded in the act (as in type II), also have a social meaning. The boundary between the two classes is of course imprecise and the same act may, under different circumstances, fall under a different class. Chopping wood, as we have already discussed, may be a functional act or a conscious distraction from work.

If contemporary sociologists insist on restricting the horizons of their empirical research and social theorising to those actions that are socially meaningful (roughly those that fall under the first class outlined above) then they will be blind to what is certainly the more common type of action (that which belongs to type II). It would be a rather pointless exercise to try to quantify the proportion of actions that fall under either class, or the percentage of people's time that is spent performing actions of either type. Any such attempt would be so reliant on such things as the definition of *an* act (which is notoriously difficult) as to be meaningless. What is clear, however, is that without actions of type II, life, let alone social life, would be impossible. Moreover, they are acts that are not necessarily easy to perform. Many such acts require considerable effort on behalf of the actor, an investment of time and an emotional or financial commitment. The social sciences, therefore, dare not take their performance for granted and should treat them as just as important and problematic as acts of type I.

As I have previously noted, one of the issues that Campbell views as central to a theory of action concerns the question of how action is possible at all, and what are the necessary conditions for the execution of actions. Now it goes without saying that many of those necessary conditions fall under what Giddens calls enabling and constraining conditions (cultural capital, the distribution of resources, the pre-existence of language and social institutions, etc.) but just as important are the antecedent internal states of the actor. To state what should be blatantly obvious to anyone that has not been taken under the spell of social situationalism, the only way to investigate a person's antecedent internal states is to ask the person involved. What should also be obvious from the previous chapters is that this is a methodological truism that transcends ontological disputes about the causal efficacy of mental states. That said, however, one's ontological position regarding the causal efficacy of PE does determine how one treats the answers one receives. An OMRist (who, it will be remembered believes in the causal efficacy of PE) is likely to treat an agent's account, when their account cites antecedent PEs as the cause of their action, as a sufficient explanation for the proximate cause of the behaviour in question. To provide a complete explanation of an individual's behaviour, of course, the OMRist will want to delve deeper into the individual's past, as well as their social and material background, but will be content that they have identified the proximate cause of behaviour once they have an explanation that cites the individual's antecedent PEs (which includes such things as their cognitive and conative states). Those persuaded by epiphenomenalism will have a similar interest in the individual's background, the social and environmental circumstances, etc., but they will be reticent about treating an individual's antecedent PEs (which for the OMRist constitutes the proximate cause) as a sufficient explanation. The reason, to summarise material from the preceding chapters, is that epiphenomenalists do not view those physical states that are referred to[2] in individuals' accounts of their actions as being sufficient causal explanations. The reasons for this are twofold. Firstly, as Dennett observes, it is all too easy to imagine that accounts of actions mirror the thought processes that are the putative cause of the action, rather than being highly edited versions. Secondly, during our discussion of supervenience we agreed with Kim that the subvenient bases of mental states have disjunctive causal powers. This means that each time a given mental state is instantiated it may be realised by a physical state with different causal powers. The same goes for actors' accounts of their actions. Even more problematic is the fact that mental states, such as *a* reason, *a* belief, *a* desire, etc., do not occur in isolation. Almost all decisions and actions, therefore, are the outcome of the combined causal powers of the physical instantiations of many different reasons, beliefs and desires. Moreover, not only are the causal powers of the physical instantiations of each of the elements of this composite state disjunctive, but their causal powers are modified by their interaction with other elements of the composite state. This latter point means only that, for example, a desire to have a glass of wine with dinner will have different causal powers dependent upon whatever other beliefs and desires the individual has. Thus if a person believes that they will have to drive home after dinner, or do some work, their belief may influence their decision about whether or not to have a drink.

From an epiphenomenalist perspective, because of the disjunctive causal powers of the subvenient base of mental states, explanations that cite mental

[2] I should remind the reader that by 'referred to' I do not mean that the individual is indicating that they have any knowledge of their causally efficacious physical states. By refers to I mean that if we apply the sort of theory of reference discussed in chapter 4 to the folk psychological terms (such as propositional attitudes) that appear in individual's accounts of their actions, then these terms can be seen as referring to the causally efficacious subvenient base not the PE.

states can only ever be partial explanations. The upshot of all this is that if one adopts an epiphenomenalist perspective towards action then one will be far more interested than OMRists in the background of the individual as well as sociobiological and neurological explanations which may usefully supplement an individual's account of the proximate cause of their behaviour. It may well be the case that OMRists will also have recourse, on occasion, to sociobiological or neurological explanations (though there is little evidence to suggest that contemporary sociologists pay anything more than lip service to the natural sciences) but when they do it will be to provide a structuring rather than triggering or proximate cause of the behaviour in question (for an explanation of this distinction see my discussion of Dretske's dual explanandum strategy in chapter 2).

Conclusion

We have covered a lot of ground over the course of this thesis and come to some very controversial conclusions. I now want to sum up and conclude by applying some of these conclusions to a real case. The example I propose to consider is that of having an animal put to sleep. Though this is perhaps not a standard sociological example, it is ideally suited to our purposes since it allows us to consider just about every aspect of action that might be of interest to social scientists.[3] It is an action where both cognitive and conative states play important causal roles; there is a central place for expert knowledge and rationality; and it can be analysed as a social action but could also plausible be construed as simply action. Moreover it is well suited to demonstrating the failure of social

[3] For those readers that would prefer a more standard sociological example, the case could easily be applied to, for example, situations where carers are involved in a person's decision to end their life due to a terminal illness.

situationalism since the situationalist perspective ignores intrapersonal processes which our previous discussion would suggest are of central importance for both an understanding and a causal explanation of the action. Furthermore it is not clear who is performing the action (vet or owner), whether it is a single action, a compound action or several different actions. In short, it provides us with ample scope to apply Campbell's critique of social situationalism to a real case and allows us to consider how a sociology that has embraced epiphenomenalism might begin to approach action and social action. We will begin by describing the case from an OMRist perspective. This will then allow us to contrast the orthodox account with a description and causal analysis of the action from both an epiphenomenalist and a social situationalist perspective. In what follows we will focus on the cause of the action as well as its execution and consider both from the two OMRist perspectives that we encountered in chapters 7 and 8.

Option 1, free action performed by the self: the decision

Acting on the advice of the veterinary surgeon an irreducible self weighed up the diagnosis, the best interest of the dog, their duty of care to the animal, the emotional consequences of the action (feelings of loss), and inaction (feelings of guilt at not having acted in the dog's best interest), and concluded that the best course of action would be to have the dog put to sleep.

Option 1: the performance

During the decision making process and the events that followed, the owner had to exercise their free will. Their decision was an act of true libertarian free will and following their decision they had to choose to continue with the course of action they had decided upon. At any point they could have freely chosen to cancel the appointment and allow their dog to die naturally. Performing the meta-action of having their dog put to sleep involved the performance of a number of more basic actions. Some of these actions would, under normal circumstances, constitute mere behaviour or habitual actions. Driving to the veterinary surgery, for example, becomes a series of actions in their own right. The emotionally charged situation means that even such things as keeping up a constant speed, signalling a turn or changing gear, are actions which require close monitoring and conscious acts of will.

Once in the waiting room part of the owner's attention is turned to maintaining a dignified presentation of self. Again, until the completion of the action the owner has to continue to freely choose to pursue the course of action set in motion by their prior decision.

Option 2, action follows from causal necessity: the decision

When the owner receives the diagnosis by telephone they come to believe that their dog has a terminal cancer and that their quality of life is unlikely to improve. Together with pre-existing beliefs and desires (such as the belief that when one becomes a dog owner one tacitly accepts that one has a duty of care towards the animal, a desire not to see their dog in pain, and so on) their newly acquired belief completes a set which provides causally sufficient conditions (*ceteris paribus*) for their decision to have their dog put to sleep. That is to say, barring the intervention of external factors the belief-desire combination will cause them to make and keep an appointment to have their dog put to sleep.

Option 2: the performance

Once made the decision sets in motion a causal chain that involves the use of practical reason. The owner's faculty for practical reason takes the decision as its input and generates a plan of action for how best to enact the decision. Its output includes such things as the making of the appointment, planning the route to the surgery, and so on. In contrast to option 1, the owner (the owner's 'self') in this scenario is passive. All the causal work is being done by their beliefs and desires, their faculty for practical reason, etc. Even behaviour that seemingly relates directly to the self, such as the desire to maintain a dignified presentation of self in the surgery, follows out of causal necessity from their pre-existing desire not to draw attention to themselves or to publicly display their emotions.

It is important to note that on both options 1 and 2 it is the owner's PEs that cause him to embark on his chosen course of action. How PE relates to neurophysiology is, of course, a matter of debate within the OMRist camp. Nevertheless, as we saw in chapters 2 and 3, any theory that deserves to be called a theory of mental causation has PE performing an ineliminable causal role. It is also important to note that on both options OMRists believe that the owner's accounts of their action refer directly (or indirectly if mediated by memory) to the causes of their action. That is to say, if the owner were to explain his action by saying that he thought 'it was the right thing to do', the referent of this account is his feeling disposed towards the proposition[4] 'having the dog put to sleep is the right thing to do' when he considers the issue.

[4] See my discussion of beliefs in ch.4 for an explanation of how 'feeling disposed towards a proposition' constitutes the referent of propositional attitudes.

An epiphenomenalist critique of the orthodox accounts

In the preceding chapters we have seen how an epiphenomenalist perspective is forced to deny the central theses of both orthodox accounts. In the first account we are forced to deny that the action was caused by the decision of an irreducible self and in the second we have to deny that the action was caused by an antecedent set of beliefs and desires. Nevertheless, there are several elements of the above accounts that are consistent with epiphenomenalism. Both accounts correctly cite internal states in the owner's mind as the cause of his actions and both accounts correctly cite the content of those mental states as being causally efficacious in the decision making process. Epiphenomenalism, as we have seen, need not deny that such things as the owner's desire not to see their dog in pain or the belief that as an owner one adopts a duty of care towards one's animals, are causally efficacious. Where epiphenomenalism and the orthodox stories diverge is that epiphenomenalists view all of the phenomenal states experienced by the owner as entirely inefficacious. Where an epiphenomenalist cites internal motivating factors (such as the desire not to see their pet in pain) they refer not to the phenomenal states themselves but to whatever neural states instantiated the content, 'I don't want to see my dog in pain'. In making this claim epiphenomenalism faces methodological and ontological problems. They face a methodological problem deriving from the fact that they have no direct access to the neurophysiological events that they claim are the real causes of the action. This lack of direct access means that they are forced to infer the existence of these states based on observed behaviour and people's accounts of their actions. In this sense epiphenomenalists are in a similar position to the PFPists considered in chapter 4. Both epiphenomenalists and PFPists have to hypothesise the existence of internal states (non-conscious beliefs and desires in the case of PFP) to account for what appears to be purposeful behaviour. Epiphenomenalism, however, also faces a serious ontological problem relating to its hypothesising the existence and causal efficacy of propositional attitudes defined in terms of their content. In chapter 4 we exploited the supervenience relation to support the claim that where there is a conscious experience of using a propositional attitude term, or holding a propositional attitude, there must be a neurally realised subvenient base which instantiates the same content as the conscious experience. From the empirical observation that people tend to do what they say they are going to do, and act in accordance with their stated beliefs and desires, we inferred that the aforementioned subvenient bases must be causally efficacious. The ontological problem for epiphenomenalism relates to the first step in this argument. Strictly speaking, as we have previously noted, epiphenomenalists cannot appeal to the supervenience relation to hypothesise the existence of a subvenient base for occurrent supervenient properties such as the PE of using propositional attitude terms or holding propositional attitudes. The reason, to repeat, is that if epiphenomenalism is true we have no means of identifying when and what a person (including oneself) is experiencing. This seriously undermines the basis from which we inferred the existence of neurally realised propositional attitudes defined in terms of their content. It ought to be possible to construct a methodology that inferred the existence of neurally realised functional states based solely on observed behaviour, including accounts of action. As the historical failures of behaviourism have shown, however, any such methodology would be severely restricted in its explanatory powers.

Another serious problem for epiphenomenalism is that the content of mental states is often linked to phenomenal states. Epiphenomenalists can happily maintain that there was some neural state that instantiated the content 'belief that the dog has terminal cancer'. Moreover, as we saw when discussing Dretske's dual explanandum strategy, they can happily adopt the functionalist strategy of claiming that the neural state gains its efficacy in virtue of its content. Epiphenomenalism, however, does not seem able to offer an adequate explanation for why mental content is causally efficacious when that content relates to phenomenal states. Suppose that the dog owner in our example was unable to go through with their decision to have their dog put to sleep. Suppose also that their reason was that they have had pets put to sleep in the past and it was an experience that they knew they would not be able to cope with again. The epiphenomenalist line would have to be that they had a neural state which instantiated the content, 'I have had pets put to sleep before and I know that I will not be able to cope with it again.' Such an explanation does not seem to be open to the epiphenomenalist, however, since the neural state in question has a phenomenal state as its referent. That is to say, it seems to imply that their previous experiences (the PE of their sense of loss and grief) were causally efficacious and that thereafter their previous *experiences* were remembered and had a causal influence on the decision making process.

The upshot of all this is that an epiphenomenalist methodology, if it is to remain true to its ontological foundations, would have to adhere to strict logical behaviourist methods. Logical behaviourism, however, is just as unsatisfactory since it too seems unable to cope with accounts such as, 'I have had pets put to sleep before and I know that I will not be able to cope with it again.' There inevitably seems to be some ‘surplus meaning’ that cannot be accounted for in terms of the observations used to determine the appropriate use of the concepts involved. There is a ‘surplus meaning’ that relates to the experiences of the dog owner that cannot be accounted for by observations of their own behaviour or the behaviour of others. Epiphenomenalism may accept the existence of this ‘surplus meaning’, but what good is that if it does not allow the experiences to have any causal role? This is no trivial problem that can be ignored as a mere anomaly. Much of human behaviour is orientated towards the pursuit of pleasure and the avoidance of pain and explanations for action frequently cite previous experiences, both pleasant and unpleasant, as the cause of present behaviour and future plans. Unless some solution can be found for this problem epiphenomenalism does not stand much of a chance of being taken seriously (particularly as the ontological foundations for a social scientific methodology).

The failure of social situationalism[5]

Despite the failures of epiphenomenalism it still has a far better chance of explaining the cause of actions than social situationalism. We have already discussed the failures of this paradigm but it is worth highlighting these failures by applying the situationalists’ methodology to the case outlined above. Doing so will highlight situationalism's inability to move beyond descriptive analysis (and flawed description at that) to genuine causal explanation. The situationalists’

[5] Because the term social situationalism denotes a set of taken for granted assumptions rather than a well formulated ‘ism’, it is impossible to predict how a social situationalist would approach the study of any topic area, particularly because there are a plethora of different sociological schools that come under the broad heading of social situationalism. In one sense, therefore, I am doing something of a disservice to many of those in the discipline by suggesting how they might tackle the topic under consideration. It could be argued that I have deliberately misrepresented the paradigm in an attempt to show its inadequacies. Nevertheless, by restricting my attention to the broad themes discussed above and applying them to a specific topic, I think the inadequacies of social situationalism can best be highlighted.

refusal to acknowledge the casual role of intrapersonal processes means that neither of the two orthodox approaches, nor the epiphenomenalist approach considered above, is open to the situationalist researcher. Rather, situationalists are far more likely to be interested in the intersubjective creation of meaning. As such situationalists might focus their attention on such things as the performance of the roles of caring pet owner and compassionate vet. As we shall see, they do not seem to have the methodological tools to provide any meaningful account of the decision, and they are equally unable to explain the private emotions and coping mechanisms that are employed by the owner after they have received the diagnosis.

The decision

The decision is surely the most important element that needs to be understood if a causal explanation of the meta-act is to be achieved and it is here that the shortcomings of social situationalism are most apparent. Where epiphenomenalists and OMRists would approach the decision by seeking to understand the internal/intrapersonal processes of the dog owner, social situationalists simply do not have the methodological tools to investigate the decision making process. With their emphasis on the interpersonal construction of meaning social situationalists are likely to focus their attention on such things as how the diagnosis is delivered to the owner (does the veterinary surgeon use euphemisms for terminal cancer, for example), how they broach the sensitive subject of whether the dog should be put to sleep, and so on. Similarly, their focus on the owner is likely to be related to such things as how they display that they understand the diagnosis that has been presented to them or how they elicit a professional opinion about what the 'right thing to do' would be, rather than how the diagnosis affects the decision making process.

It is by no means inevitable that the owner of a dog diagnosed with terminal cancer will choose to have their pet put to sleep. It is quite possible that they will find the experience too harrowing to go through with or they might have a moral objection to having animals destroyed. As such it seems that the only way to causally explain the owner's decision is by looking towards their internal states. The way that the diagnosis is delivered, or how the owner and vet interact during the delivery of the diagnosis, are irrelevant to the decision making process. It is their having an internal state (a physical state for epiphenomenalists and a conscious mental state for OMRists) with the content 'my dog has terminal cancer', and how that state relates to the rest of their beliefs and desires, that is important. Since situationalists typically view such states as inaccessible (or worse non-existent) they cannot be used by situationalists as an explanatory resource. This means that the decision is beyond their remit and consequently the meta-act is rendered unintelligible.

The performance

Part of the performance of the meta-action takes place in a public arena and involves interpersonal communication, so here at least the situationalists have some material with which they can work. Nevertheless, the interactional element is of secondary importance if a causal account is the goal of the research and, perhaps more importantly, the situationalists methodological approach to interaction means that a descriptive analysis is all that they can deliver. Let's begin by breaking up the performance into three distinct phases: there is the time between the owner making the appointment to have their dog put to sleep and their embarking on the journey to the veterinary surgery, there is the journey itself, and there is the time spent in the surgery. If we suppose that the first and second stages were performed in private then there is little material available for situationalists to analyse. There is no interaction, no interpersonal construction of meaning, and (arguably) no roles being performed. Nevertheless, from both an OMRist and an epiphenomenalist perspective there is a lot going on that stands in need of causal analysis. To list few elements that might be of interest to researchers, there is the question of how the owner copes psychologically with the knowledge that in a few hours they are going to have to drive to the veterinary surgery and face what they know is going to be an extremely harrowing experience. At any time they could freely choose (according to the OMRist camp) to avoid the trauma by cancelling the appointment, or (according to epiphenomenalists) their brain might instantiate the content 'I just can't go through with it'. They will also face the non-trivial task of ensuring their dog has the opportunity to go to the toilet, of 'saying goodbye' in a private space and before they face the public arena of the surgery. There are the emotions of grief, love, dread and guilt, that, despite viewing the emotions from two different perspectives (one physical/chemical and the other in terms of their PE), both OMRists and epiphenomenalists agree causally influence the owner's actions. All this is beyond the scope of situationalism.

Jumping to the final stage when the owner enters the surgery, situationalists will be interested in how the owner presents themselves to the vet and members of the public. They might want to investigate such things as how the owner manages to present a dignified appearance (or fails to present a dignified appearance) amongst people who are attending the surgery for routine vaccinations and check-ups. Or how the vet performs the role of the caring professional that is trying to present the image that they 'understand' what the owner is going through and believe it is 'the right thing to do'. They may also be interested in the Goffmanesque designation of 'front' and 'back' regions and are sure to be fascinated by the fact that the owner is forced to sit in the public waiting room but offered the opportunity to exit through a back door. I would not wish to deny that an analysis of such factors is essential for an understanding of how the interaction progresses. Nor would I wish to suggest that they are irrelevant to a causal understanding of the act (knowing that one will be treated with compassion and spared the embarrassment of exiting through the waiting room in floods of tears no doubt has a causal impact on, for example, whether one attends the surgery or calls the vet out). As soon as these elements are treated as 'performances', however, their explanatory power is severely diminished. Unless observed behaviour (the owner's grief or the vet's compassion) is treated as displaying real psychological states (and remember that both the OMRist camp and the epiphenomenalist camp do treat these displays as reflecting real psychological states) the responses they elicit cannot be treated as genuine. There is surely something missing in an account of the owner's tears if they are treated as analogous to an atheist's swearing an oath on the bible in court. What is missing is not some pointless discussion of 'private' and unobservable internal states. What is missing is the essence of action and the key to causal explanation.

Concluding remarks

The *sui generis* real nature of subjective experience was the motivating factor behind our refusal to accept reductionist accounts of mental causation and was also cited as the defining characteristic of selfhood. Paradoxically the irreducibility of PE provides both the ontological foundations for epiphenomenalism and its greatest challenge. Throughout this thesis we have been confronted by the ontological and methodological problem that if epiphenomenalism is true we should not be able to discuss the nature of PE. If epiphenomenalism is true, any discussion of PE has to be viewed as merely the end product of complicated neuronal processes together with the interaction between those processes and Worlds 1 and 3. As such everything we have said ‘about’ PE has no causal connection to PE. This is likely to be seen by many as not only undermining the foundations of just about every argument presented in this thesis, but as a complete refutation of epiphenomenalism. I have to confess that I can see no solution to this problem.

Despite its inadequacies epiphenomenalism does deserve to be taken seriously and now seems to be the right time to get the discussion started. Over the last fifty years there has been a growing consensus within the philosophy of mind that PE is not reducible to physical states. This consensus has not, however, led to the resurgence of interest in dualism that might have been expected. The overwhelming majority of contemporary theories of mind may be characterised by an adherence to the three principles set out in the introduction plus the belief in mental causation. There seems to be something fundamentally inconsistent in this set (or at least in the formulation of the principles) and sooner or later something will have to give. That there is some inconsistency is now, finally, beginning to be accepted. Of the four principles, however, P1 (the irreducibility of mental states) is the only one to have received serious and sustained attention. Here there have been two main strategies: some have attempted to provide reductive accounts and some have argued that the seeming irreducibility of mental states is just an epistemological consequence of our approach to the problem or of our limited cognitive abilities (as McGinn argues). Though all these attempts should be welcomed, there is surely much that could be learned from placing the other three principles under equal scrutiny. The principle of causal explanatory exclusion, being a logical truism, is unlikely to be rejected. The other two principles (of the causal closure of the physical and the belief in mental causation), however, are neither logical truths nor empirical certainties. If we are ever to solve the mind-body problem then both need to be subjected to the same scrutiny as has been applied to reductionist and anti reductionist arguments. Emergence (in the sense of emergent powers) and epiphenomenalism have been woefully neglected by philosophers of mind. The knowledge that we stand to gain by pushing these arguments as far as they will go should not be underestimated (regardless of whether either theory turns out to be true).

Bibliography

Archer, M.S. 2000. *Being Human: The Problem of Agency*. Cambridge: Cambridge University Press.

Archer, M.S. forthcoming. *Structure, Agency and the Inner Conversation*.

Armstrong, D.M. 1994. 'Introspection'. In Q. Cassam (ed.) *Self Knowledge*. Oxford: Oxford University Press.

Baker, L.R. 1987. 'Metaphysics and mental causation'. In J. Heil and A. Mele (eds.) q.v., 75-96.

Berkeley, I. 1997. 'Some myths of connectionism'. *URL: http://www.ucs.lousiana.edu/~isb9112/dept/phil341/myths/myths.html*.

Bhaskar, R. 1998. *The Possibility of Naturalism: A Philosophical Critique of the Contemporary Human Sciences*. London: Routledge.

Block, N. 1978. 'Troubles with functionalism'. In W. Savage (ed.) *Perception and Cognition: Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science*, IX: 261-326. Minneapolis: University of Minnesota Press.

Campbell, C. 1996. *The Myth of Social Action*. Cambridge: Cambridge University Press.

Campbell, D. 1987. 'Evolutionary epistemology'. In G. Radnitzky and W. Bartley (eds.) *Evolutionary Epistemology, Rationality and the Sociology of Knowledge*. Chicago: Open Court: 47-89.

Campbell, K. 1970. *Body and Mind*. London: The MacMillan Press.

Chalmers, D. 1996. *the Conscious Mind: In Search of A Fundamental Theory*. Oxford: Oxford University Press.

Churchland, P.M. and Churchland, P.S. 1998. *On The Contrary: Critical Essays, 1987-1997*. London: The MIT Press.

Churchland, P.M. 1998. 'Folk Psychology'. In P.S. Churchland and P.M. Churchland (eds.) q.v., 3-16.

Clark, T.W. 1998. 'Function and phenomenology: closing the explanatory gap'. In J. Shear (ed.) *Explaining Consciousness: The Hard Problem*. Cambridge, MA: A Bradford Book: 45-60.

Cohen, J.L. 1996. 'Does belief exist?'. In A. Clark and P. Millican (eds.) *Connectionism, Concepts and Folk Psychology*. Oxford: Clarendon Press: 265-76.

Damasio, A.R. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: G. P. Putnam.

Davidson, D. 1980a. *Essays on Actions and Events*. Oxford: Oxford University Press.

Davidson, D. 1980b. 'Mental events'. In D. Davidson, q.v., 207-28.

Davidson, D., 1980c. 'The material mind'. In D. Davidson, q.v., 245-60.

Davidson, D. 1993. 'Thinking causes'. In J. Heil and A. Mele (eds.) q.v., 3-17.

Dennett, D. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: A Bradford Book.

Dennett, D. 1991a. 'Two contrasts: folk craft versus folk science, and belief versus opinion'. In J. Greenwood (ed.) *The Future of Folk Psychology*. Cambridge: Cambridge University Press: 135-48.

Dennett, D. 1991b. 'Ways of establishing harmony'. In B. McLaughlin (ed.) q.v., 119-30.

Dennett, D. 1993. *Consciousness Explained*. Harmondsworth: Penguin Books.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes.* Cambridge, MA: A Bradford Book.

Dretske, F. 1991. 'Dretske's replies'. In B. McLaughlin (ed.) q.v., 180-221.

Elster, J. 1984. *Ulysses and the Sirens: Studies in Rationality and Irrationality.* Cambridge: Cambridge University Press.

Fodor, J. 1968. 'The appeal to tacit knowledge in psychological explanation'. *Journal of Philosophy*, 65: 627-40.

Fodor, J. 1990. 'Making mind matter more'. In J. Fodor *A Theory of Content and Other Essays*. Cambridge, MA: A Bradford Book.

Fodor, J.A. and Pylyshyn, Z. W., 1988. 'Connectionism and cognitive architecture: A critical analysis'. *Cognition*, 28: 3-71.

Giddens, A. 1991. *Modernity and Self Identity: Self and Society in The Late Modern Age*. Cambridge: Polity Press.

Goffman, E. 1971. *The Presentation of Self in Everyday Life*. Harmondsworth: Penguin.

Gould, S. and Lewontin, R. 1979. 'The Spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme'. *Proceedings of the Royal Society of London*, 205 B: 581-98.

Greenwood, J. 1991a. *The Future of Folk Psychology: Intentionality and Cognitive Science*. Cambridge: Cambridge University Press.

Greenwood, J. 1991b. 'Reasons to believe'. In J. Greenwood (ed.) *The Future of Folk Psychology: Intentionality and Cognitive Science*. Cambridge: Cambridge University Press: 70-92.

Grush, R. and Churchland, P.S. 1998. 'Gaps in Penrose's toilings'. In P.M Churchland and P.S. Churchland (eds.) *On the Contrary: Critical Essay, 1987-1997*. Cambridge, MA: A Bradford Book: 205-29.

Hare, R.M. 1952. *The Language of Morals*. Oxford: Clarendon Press.

Haskar, W. 1999. *The Emergent Self*. London: Comell University Press.

Hebb, D.O. 1949. *The Organization of Behavior; A Neuropsychological Theory*. New York: Wiley.

Heil, J. and Mele, A. (eds.) 1993. *Mental Causation*. Oxford: Clarendon Press.

Hempel, C. 1965. 'Studies in the logic of scientific explanation'. *Philosophy of Science*, 15: 135-75.

Hofstadter, D.R. 1979. *Godel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books.

Holland, J. 1998. *Emergence: From Chaos to Order*. Oxford: Oxford University Press.

Honderich, T. 1988. *A Theory of Determinism: The Mind, Neuroscience and Life-Hopes*. Oxford: Clarendon Press.

Horgan, T. 1989. 'Mental quausation'. *Philosophical Perspectives*, 3: 47-76.

Horgan, T. 1991. 'Actions, reasons, and the explanatory role of content'. In B. McLaughlin (ed.) q.v., 73-101.

Horgan, T. 1993. 'From supervenience to superdupervenience: meeting the demands of a material world'. *Mind*, 102: 555-86.

Horgan, T. and Woodward, J. 1991. 'Folk psychology is here to stay'. In J. Greenwood (ed.) *The Future of Folk Psychology*. Cambridge: Cambridge University Press.

Jackson, F. 1982. 'Epiphenomenal qualia'. *Philosophical Quarterly*, 32: 127-136.

Kauffman, S. 2000. *Investigations*. Oxford: Oxford University Press.

Kim, J. 1984a. 'Concepts of supervenience'. *Philosophy and Phenomenological Research*, 65: 153-176.

Kim, J. 1984b. 'Epiphenomenal and supervenient causation'. In J. Kim (1993f) q.v., 92-108.

Kim, J. 1991. 'Dretske on how reasons explain behavior'. In B. McLaughlin (ed.) q.v., 52-72.

Kim, J. 1993a. 'Can supervenience and "non-strict laws" save Anomalous Monism'. In J. Heil and A. Mele (eds.) q.v., 19-26.

Kim, J. 1993b. 'Mechanism, purpose, and explanatory exclusion'. In J. Kim (1993f) q.v., 237-264.

Kim, J. 1993c. 'The myth of nonreductive materialism'. In Kim (1993f) q.v., 265-283.

Kim, J. 1993d. 'The nonreductivist's troubles with mental causation'. In J. Kim (1993f) q.v., 336-57.

Kim, J. 1993e. 'Postscripts on mental causation'. In J. Kim (1993f) q.v., 358-367.

Kim, J. 1993f. *Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press.

Kim, J. 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: A Bradford Book.

LePore, E., and Loewer, B. 1989. 'More on making mind matter more'. *Philosophical Topics*, 17: 175-91.

Lewis, D. 1970. 'How to define theoretical terms'. *Journal of Philosophy*, 67: 427-46.

Lewis, D. 1972. 'Psychophysical and theoretical identifications'. *Australasian Journal of Philosophy*, 50: 249-58.

Libet, B. 1993. 'The neural time factor in conscious and unconscious events'. *Experimental and Theoretical Studies of Consciousness*, 174: 123-146.

Libet, B., Gleason, C., Wright, D. and Pearl, D. 1983. 'Time of conscious intention to act in relation to onset of cerebral activity (readiness potential)'. *Brain,* 106: 623-642.

Marcuse, H. 1998. *Eros and Civilization: A Philosophical Inquiry into Freud.* London: Routledge.

McLaughlin, B. (ed.) 1991. *Dretske and His Critics.* Oxford: Basil Blackwell.

McLaughlin, B. 1993. 'On Davidson's response to the charge of epiphenomenalism'. In J. Heil and A. Mele (eds.) q.v., 27-40.

Mead, G. 1934. *Mind, Self, and Society: From the Standpoint of a Social Behaviorist.* London: The University of Chicago Press.

Nagel, T. 1995. 'What is it like to be a bat'. In W. Lyons (ed.) *Modern Philosophy of Mind.* London: Everyman: 159-74.

Penrose, R. 1994. *Shadows in the Mind: A Search for the Missing Science of Consciousness.* Oxford: Oxford University Press.

Popper, K. 1972a. 'Epistemology without a knowing subject'. In K. Popper *Objective Knowledge: An Evolutionary Approach.* Oxford: Clarendon Press: 106-52.

Popper, K. 1972b. 'Evolution and the tree of knowledge'. In K. Popper *Objective Knowledge: An Evolutionary Approach.* Oxford: Clarendon Press: 256-84.

Popper, K. 1982. *The Open Universe: An Argument for Indeterminism.* London: Routledge.

Popper, K. 1987. 'Natural selection and the emergence of mind'. In G. Radnitzky and W. Bartley (eds.) *Evolutionary Epistemology, Rationality and the Sociology of Knowledge*. Chicago: Open Court: 139-55

Popper, K. 1994. *The Myth of the Framework*. London: Routledge.

Popper, K. and Eccles, J. 1977. *The Self and Its Brain*. London: Springer International.

Putnam, H. 1975. 'The meaning of "meaning"'. In K. Gunderson (ed.) *Language Mind and Knowledge, vol. 7, Minnesota Studies in the Philosophy of Science*, Minnesota: University of Minnesota Press.

Ramsey, W., Stich, S. and Garon, J. 1996. 'Connectionism, eliminativism, and the future of folk psychology'. In S. Stick (ed.) q.v., 91-114.

Rorty, R. 1989. *Contingency, Irony, and Solidarity*. Cambridge: Cambridge University Press.

Rose, S. 1992. *The Making of Memory*. London: Bantam Press.

Ryle, G. 1949. *The Concept of Mind*. Middlesex: Penguin Books.

Searle, J. 1992. *The Rediscovery of the Mind*. London: A Bradford Book.

Searle, J. 2000. 'Consciousness, free action and the brain'. *Journal of Consciousness Studies*, 7: 3-22.

Searle, J. 2001. *Rationality in Action*. Cambridge, MA: A Bradford Book.

Smart, J.J.C. 1981. 'Physicalism and emergence'. *Neuroscience*, 6: 109-113.

Sosa, E. 1984. 'Mind-body interaction and supervenient causation'. *Midwest Studies in Philosophy*, 9: 271-81.

Spence, S. 1996. 'Free will in the light of neuropsychiatry'. *Philosophy, Psychiatry, & Psychology*, 3: 75-90.

Sperry, R.W. 1980. 'Mind-brain interactionism: mentalism, yes; dualism, no'. *Neuroscience*, 5: 195-206.

Sperry, R.W. 1991. 'In defence of mentalism and emergent interaction'. *Journal of Mind and Behavior*, 12: 221-45.

Stich, S. (ed.) 1996. *Deconstructing the Mind.* Oxford: Oxford University Press

Stich, S., and Nichols, S., 1996. 'How do minds understand minds? Mental simulation versus tacit theory'. In S. Stick (ed.) q.v., 136-67.

Stich, S. and Ravenscroft, I. 1996. 'What Is folk psychology'. In S. Stich (ed.) q.v., 115-135.

Stich, S. 1996. 'Deconstructing the mind'. In S. Stich (ed.) q.v., 3-90.

Trigg, R. 1993. *Rationality and Science: Can Science Explain Everything?* Oxford: Blackwell.

van Gelder, T. 1991. 'What is the "D" in "PDP"? A survey of the concept of distribution'. In W. Ramsey, S. Stich, and D. Rumelhart, (eds.) *Philosophy and Connectionist Theory.* Hillsdale, N.J.: Lawrence Erlbaum: 33-60.

Velmans, M. 1991. 'Is human information processing conscious'. *Behavioural and Brain Sciences*, 14: 651-746.

Vrancken, M. 1989. 'Schools of thought on pain'. *Social Sciences and Medicine*, 3: 435-44.

Webster, R. 1995. *Why Freud Was Wrong: Sin, Science and Psychoanalysis.* London: Harper Collins.

Wittgenstein, L. 1968. *Philosophical Investigations.* Oxford: Blackwell.