Journal of Consciousness Studies
www.imprint-academic.com/jcs

Anthony Rudd

Phenomenal Judgment and Mental Causation

Abstract: *This paper defends and develops an argument against epiphenomenalism, broadly construed. I argue first for a definition of epiphenomenalism which includes 'non-reductive' materialism as well as classical dualistic epiphenomenalism. I then present an argument that if epiphenomenalism were true it would be impossible to know about or even refer to our conscious states — and therefore impossible even to formulate epiphenomenalism. David Chalmers has defended epiphenomenalism against such arguments; I consider this defence and attempt to show that it fails. I conclude that an adequate account of mental causation requires us to abandon the principle of the causal closure of the physical, and attempt to rebut charges that it would be 'unscientific' to do so.*

In this paper I shall consider, and endorse, an argument that epiphenomenalism is, if not exactly self-contradictory, at least self-defeating. I shall however be using 'epiphenomenalism' in a rather wider sense than is usual; this will of course mean that the argument, if applicable to epiphenomenalism in this wide sense, is itself of greater significance than might have been thought.

I

I shall start by explaining my usage. Classic Epiphenomenalism is a kind of dualism, according to which mental events are really different from physical events, are caused by them, but have no causal powers themselves. My usage is wider; it includes not only Classic Epiphenomenalism, but other positions that are commonly distinguished from it. The reason why I adopt the wider usage is, of course, that I think the differences between the various positions are in fact relatively trivial, and that the argument which I am going to consider, though formulated with Classic Epiphenomenalism in mind, actually applies to a much wider range of philosophical views.

In my sense, then, 'epiphenomenalism' is the conjunction of the following assertions:

Correspondence: Dr A.J. Rudd, Department of Philosophy, University of Hertfordshire, Watford Campus, Wall Hall, Aldenham, Herts WD2 8AT, UK.

(1) There are irreducibly real conscious states (events, processes etc.).
(2) They have no causal powers.

The first claim is simply that people really (ultimately, in the last analysis, etc.) do have conscious feelings, sensations, thoughts, internal monologues, perceptual experiences etc.; to talk about such things isn't just a perhaps useful but ultimately dispensable (at least in theory) way of referring to a bottom-line reality consisting of functionally organized brain states. In other words it is simply the denial of eliminative or reductive materialism. I shall assume in what follows that (1) is correct. It does not necessarily commit one to dualism, for it would be accepted by 'non-reductive' materialists and dual aspect theorists, who would argue that the conscious states are identical with certain physical states. They would however deny that the purely physical descriptions of those states are sufficient to fully characterize what they are.

As for (2), my claim is that it follows from

(3) The physical world is a closed causal system.

(3) is not necessarily determinism, though of course it is compatible with it. The point is not to deny that there may be purely random uncaused events, or that causation may be probabilistic rather than strictly deterministic. But (3) does insist that the only causes of physical events are other physical events. Insofar as an event in the physical world is causally explicable at all, it is explicable in terms of purely physical events, processes and laws. If this is the case, then it seems there can be no room left over for mentalistic explanations of physical events. Hence the claim that (2) follows from (3) — and (3) is certainly widely accepted. It would hardly seem avoidable for physicalists,[1] but is accepted even by many who are not physicalists. Frank Jackson has suggested that to deny it would be to sound 'like someone who believes in fairies' (Jackson,1990, p. 470). Despite that risk I shall argue that, if (2) is directly implied by (3), and if the conjunction of (2) with an obvious truth (which, although there are eminent philosophers who would deny it, I take (1) to be) produces an absurdity, then (2) — and therefore (3) — must be false.

I have claimed that non-reductive materialists are committed to asserting (2) because it follows from (3). Various attempts have been made to deny the implication from (3) to (2), however. Davidson presented it as one of the virtues of his version of non-reductive materialism that it enabled him to reconcile the causal closure of the physical world with the reality of the causal role played by mental events (see Davidson, 1980). Mental events have physical effects because those very events are (under a different description) physical events. But of course, it is a standard objection to Davidson's account that it does not really avoid epiphenomenalism. (Taken in something like my sense.) On Davidson's account, the events that are mental events have a causal efficacy in the physical world; but they have that efficacy only in virtue of their also being physical events. So it is not *qua* mental that they have causal powers, and therefore their mental aspects are not causally relevant. If those events had only their physical characteristics, and no mental aspects at all, it would make no

[1] To abandon it would be to abandon anything that was worth calling 'physicalism'.

difference to their causal efficacy. So mentality (or consciousness) remains, on Davidson's account, merely epiphenomenal.

Davidson has responded to this charge by appealing to the following principle:

(4) Causation is a pure extensional relation between particular events as such, rather than between events qua bearers of certain properties.

Hence, he argues, one cannot in fact say what his critics think he is committed to saying — that it is in virtue of its physical rather than its mental properties that an event is causally effective (see Davidson, 1993). Now, even if one accepts (4), it is not clear that it does yield an adequate response to the charge of epiphenomenalism (see McLaughlin, 1993; also Hutto, 1999). But in any case, (4) is a very implausible and counter-intuitive principle; in ordinary life we constantly distinguish between causally relevant and non-relevant features of events (and indeed objects).[2] Of course Davidson has his reasons for asserting (4), but to fully consider them would involve an examination of his theory of causation, his event ontology and, ultimately, his underlying philosophical methodology. Such an examination would, I think, show that Davidson is quite deeply mistaken about all these matters; but a full account of this would involve (at least) a separate paper in itself, and I shall not try to pursue these issues further here.

If non-reductive materialists don't want to appeal to (4), I think the most plausible response they could make to the charge of epiphenomenalism would be to invoke Kripkean metaphysical necessity. If some mental event is identical with some physical event, then it is necessarily identical with it. Hence, the speculation above — that on another possible world that physical event would still occur without any mental aspect — is not really coherent. Just as the Morning Star could not fail to be the Evening Star, so this stab of pain could not fail to be the brain event that it is.

There are a range of defences that could be adopted against this Kripkean move. The first would be to criticize the very notion of metaphysical necessity, but that is not a subject I can discuss here. The second line of defence would involve accepting the Kripkean account of identity, but then following Kripke's own line by arguing that, since conscious states and neural events are obviously not *necessarily* identical, it follows that they can't be identical at all (see Kripke, 1980, pp. 144–55). Thirdly, even if one accepted that they were necessarily identical, and that therefore there couldn't be a Twin Earth physically identical to ours but with no consciousness, this still would not show that consciousness itself is causally effective; merely that it necessarily goes along with physical states that are. It would simply be a brute (and deeply mysterious) fact that certain physical (and therefore causally effective) states also had mental aspects necessarily tied to them, although it was solely in virtue of their physical aspects that they were causally effective.[3]

The point can perhaps be brought home by noting that the same argument could be run with Classic (dualistic) Epiphenomenalism. One might assert the necessary identity of the brain state that causes such and such bodily movements and the brain

---

[2] As in the well known example, originally due to Dretske; the diva's singing shattered the glass in virtue of its pitch, rather than in virtue of its semantic content.

[3] The mysterious and brute nature of this supposed connection is, of course, responsible for the suspicion that such a connection could not be necessary at all.

state that causes such and such conscious states. One could then conclude that there couldn't be a possible world which was physically indistinguishable from ours, but in which there were no conscious states. But this Kripkean neo-Classical Epiphenomenalism would still be Epiphenomenalist, in that it would still allow no actual causal efficacy to the conscious states themselves. So if one adds Kripkean semantics to either Davidson or Huxley, one may be able to avoid the possibility of a non-conscious but physically indiscernible Twin Earth, but the intuition remains that it would be avoided in the *wrong way*, one that still doesn't make consciousness anything more than a parasitic observer.

Not only that but also — this is the fourth and final line of defence — the Kripkean move does nothing at all to avoid the possibility of a Close Cousin Earth Without Consciousness. For we can imagine a possible world which is physically very like ours, the only difference being that the brains of the creatures there are very slightly, but systematically, different from ours. They look very similar and behave very similarly to us. But they might be wholly unconscious. Not just slightly different in their inner lives, but wholly lacking them. Even if it were true that creatures physically identical to us would have to have consciousness, one would only have to make a slight physical change for the argument to no longer go through. That it seems to allow for a Twin Earth Without Consciousness (or Zombie Earth, as it is sometimes described[4]) has always been one of the main intuitive sources of resistance to Epiphenomenalism. But even if Kripke can help the Epiphenomenalists to avoid this possibility, he can't help them to rule out the possibility of Close Cousin Zombie Earth, and this seems almost as counter-intuitive as the original Twin Zombie Earth.[5]

I haven't attempted to consider all the ways in which non-reductive materialism might be defended against the charge of epiphenomenalism. However, the argument that the causal closure of the physical allows no room for mental causation is intuitively very powerful and I can't see any way around it that looks at all plausible. If I am right about this, then the difference between Classic Epiphenomenalism and non-reductive materialism comes down to the difference between saying that brain states cause distinct but causally inert conscious states, and saying that those brain states have causally irrelevant conscious aspects. This difference is actually quite trivial. Though, technically, one view is dualist and the other physicalist, the distinction between them hardly marks anything that can be thought of as a great metaphysical divide. The important features of both views are the ones they share — the claims that consciousness is real but plays no active causal role. In what follows I shall consider an argument that this conjunction of claims — epiphenomenalism in my sense — is self-stultifying. Non-reductive materialists who do think they can avoid epiphenomenalism can, I hope, still read what follows with interest as a demonstration of the necessity, rather than just the desirability, of their avoiding it.

---

[4] See the symposium on 'Zombie Earth' in *Journal of Consciousness Studies*, **2** (4) (1995).

[5] This argument has some similarities to, though also differences from, Kim's critique of Davidson on supervenience. See e.g. Kim (1989).

II

The basic idea of this argument is that epiphenomenalism cannot be consistently asserted. It is a claim about our mental states, but if it were true then we would be unable to know about or talk about those mental states. So if it were true it could not be asserted. Arguments along these lines have been around for a while, but I shall focus specifically on the version presented by John Foster in his book *The Immaterial Self* (Foster, 1991, Ch. 6, sec. 5). Foster introduces the argument by stating 'I cannot see how, if epiphenomenalism were true, the mind could form a topic for overt discussion' (*ibid*., pp. 190–1). In fact Foster presents two related arguments, one about knowledge, the other about reference. The first is that 'if mental items have no causal access to our speech centres, the notion of an introspective report collapses' (*ibid*., p. 191). If the utterances we make about our (conscious) mental states are entirely caused by processes other than those states, then it cannot be supposed that those utterances express any knowledge of those states.

The second argument follows on from the first, giving it a further twist. If our conscious states have no causal influence on the utterances we suppose we make about them, then we cannot in fact suppose that those utterances *are* about the conscious states at all.

> [I]f the mental contributes nothing to the way in which the linguistic practices involving 'psychological' terms are developed and sustained in the speech-community . . . then . . . the terms should be analysed in a purely behaviourist or functionalist fashion — which would deprive the epiphenomenalist of the linguistic resources to enunciate his thesis. (*ibid*., p. 191)

There is a further possible argument, to the effect that if we understand language in a purely mechanistic, physicalist fashion, as causally unconnected to consciousness, then we lose the sense that it is expressive of meanings at all. On this account, a non-mentalistic theory of language undermines the very idea of meaning and thus, *inter alia*, renders its own assertion meaningless.[6] Foster, however, doesn't push the argument this far. His is a restricted argument about one region of language: that if we are to refer to our mental states, there must be some causal link between those states and our assertions. If this is right, then epiphenomenalism, which presupposes by its own assertion that we can refer to our conscious states, but denies the casual link, must be at any rate self-stultifying.

Foster does not claim that his argument, strictly speaking, *refutes* epiphenomenalism. It doesn't show that there is a self-contradiction in the epiphenomenalists' description of the world. But it does show that it cannot be asserted, or even privately considered, without presupposing its falsity and, as Foster says, this 'renders it the sort of position that cannot be seriously entertained' (Foster, 1991, p. 192). He then proceeds to point out that the same reasoning applies to the sort of non-reductive physicalism which I have already argued for including as a form of epiphenomenalism. As Foster notes, if his original arguments are correct then they will also show that we cannot consistently claim to know about, or even refer to, conscious *properties* or *aspects* (any more than we can to conscious *states*) while asserting their causal irrelevance (*ibid*., pp. 194–6). If Foster's argument is right then it is highly

---

[6] Norman Malcolm develops an argument along these lines in Malcolm (1968).

significant, for it will enable us to dismiss not only classic epiphenomenalism, but any version of materialism which allows that there are irreducibly conscious properties. In the following section I shall consider a recent defence of epiphenomenalism, not specifically against Foster but against arguments very like his. This is provided by David Chalmers in his book, *The Conscious Mind.*[7]

III

Chalmers, despite some qualifications, accepts that his own position of 'naturalistic dualism' is effectively epiphenomenalist;[8] and he acknowledges that this gives rise to what he calls 'The Paradox of Phenomenal Judgment'. If the universe is a closed physical system, and there are irreducibly real states of consciousness, how is it that we can form and express judgments about states of consciousness? As he says: 'It seems very strange that our experiences should be irrelevant to the explanation of why we *talk* about experiences . . . It might be claimed that [epiphenomenalism] is incompatible with our *knowledge* of experience, or with our ability to *refer* to experiences' (Chalmers, 1996, pp. 159–60). Chalmers doesn't, rather surprisingly,[9] mention Foster specifically, but clearly the arguments he sets himself to refute are essentially those that Foster presented.

Chalmers considers two possible versions of the argument that, given epiphenomenalism, I cannot know my own conscious states. One version points out that, if epiphenomenalism were correct, there could be a being physically identical to me, but without consciousness ('my zombie twin'). This twin would, *ex hypothesi* make all the same utterances that I do. But he would not be justified in what he said about having conscious states. However, if my utterances are produced by the same wholly physical mechanisms as his, then neither am I justified in my utterances about my conscious states. The second version Chalmers considers depends on a causal theory of knowledge, which would analyse knowledge as true belief which is causally connected to the object or state of affairs it is about. On this view, if conscious states are causally inert, they cannot be objects of knowledge.

Chalmers has set up these arguments so that they depend on contentious premises in epistemology. Accordingly, he deflects the arguments by rejecting the causal theory, and the idea that justification attaches to the mechanisms by which beliefs are formed. Instead, he asserts simply that 'it is having the experiences that justifies the beliefs' (Chalmers, 1996, p. 196). In other words, no mechanism is needed to mediate between a conscious experience and a belief about it: 'To have an experience is automatically to stand in some sort of intimate epistemic relation to the experience.' (*Ibid.*, pp. 196–7) I think this is right, and Chalmers uses this principle to counter the specific versions of the argument that he discusses. To have a conscious state is, in some measure, to know that one has it. So, Chalmers wants us to conclude, there can be no problem; epiphenomenalism asserts that we do have conscious states, so of course it allows that we can know them, since we can't have them without knowing them. However, this move does not help him to cope with the argument Foster

[7] Chalmers (1996), Ch. 4, secs. 4 and 5; Ch. 5.

[8] See Chalmers (1996), Ch. 4, sec. 4.

[9] Surprising since he does list Foster's book in his bibliography and refers to it elsewhere.

presents, which is slightly weaker than those that Chalmers considers, but is still strong enough to be devastating to epiphenomenalism.

Foster concedes, at least for the sake of argument, that epiphenomenalism may allow us to *know* our inner states. But, he insists, it cannot explain how we could *express* that knowledge: 'even if a subject retains an introspective knowledge of his mental states, his utterances do not count as *expressing* that knowledge if it contributes nothing to their production' (Foster, 1991, p. 191). This seems right. For me to be in pain may involve my knowing I am in pain, but it is hard to see how public speech could count as expressing this inner knowledge unless that state of being in/knowing I am in pain played some causal role in bringing about the utterance 'I am in pain'. Chalmers' counter-argument, then, does nothing to even engage with the claim that epiphenomenalism cannot be publicly articulated. It may still seem that Chalmers has at least rescued the possibility of my having private thoughts about my inner states, but it isn't even clear that he has done that. For a strict epiphenomenalism, all my conscious states are caused solely by physical events, and are themselves causally inert. So it cannot allow that one conscious state can have a causal effect on any other conscious state. But then, even if each conscious state contains within itself a knowledge that one is in that state, it is still hard to see how one can be in a single, unitary state of knowing that one is having various different conscious experiences.

This argument opens up in two different directions. The first concerns the way in which different conscious states co-exist at the same time (e.g. a stab of pain, the smell of cooking, a thought about the philosophy of mind). If each is the product of a distinct brain state, and if none has any direct connection with any of the others, then it is hard to see how they could come together as they do to constitute a single complex experience. One — desperate — response would be to follow Dennett and claim that the unity of consciousness is an illusion. Since there is no central 'Cartesian Theatre' in the brain where everything comes together physically, there can be no corresponding unity in conscious experience.[10] But the correct reply is to turn the argument around. There patently is a 'unity of apperception' as Kant puts it, and, indeed, as he correctly argued, it is a transcendentally necessary condition for there to be any experience — and therefore any illusions — at all. In which case, if there is no corresponding unity of brain processes, it follows that consciousness has a kind of unity which is not physically explicable.[11] An epiphenomenalist would not, of course, want to embrace Dennett's radical kind of materialism, and would presumably want to assert the unity of consciousness. But this is the whole problem, for it is hard to see how one could construct such a unity out of discrete conscious states, caused by distinct brain processes and with no direct causal connections existing between them.

The second problem has to do with memory. Even if the epiphenomenalist could give some account of the simultaneity of conscious states, how can she explain my having a conscious state which is a memory of a previous one? Surely this must

---

[10] This is the central argument of Dennett (1991).

[11] So the empirical data which Dennett assembles ironically supports an updated version of one of the classic Cartesian arguments for dualism — that the mind has an intrinsic unity whereas matter is always theoretically divisible.

involve some causal link between the present and the past state?[12] Chalmers tries to get round this problem by denying that memory is necessarily a causal notion (Chalmers, 1996, pp. 200–1). If the underlying brain states are causally connected, that is good enough; the conscious memory states don't need to be connected themselves. 'What is important is that a non-reductive theory can save the appearances by giving a mechanism by which true beliefs about one's past experiences are formed' (*ibid*., p. 201). Like many other theories that claim to 'save the appearances', this one gives the impression that we have been fobbed off — we are indeed left with the appearances, but the reality of memory has been lost. I find Chalmers' response pretty feeble, but I don't want to rest on an appeal to intuitions of plausibility — anyone who was strongly impressed by such things would hardly be likely to adopt epiphenomenalism in the first place. It is after all characteristically a theory of last resort — one into which people are pushed by the sense that all the alternatives are even less plausible.

I think that these arguments do put considerable pressure on the idea that epiphenomenalism can allow for even a private inner knowledge of one's own inner states, but they are not perhaps entirely decisive. What does seem clear is that, for an epiphenomenalist, one's overt utterances could have no connection with these private inner states, and therefore couldn't be counted as expressing knowledge of them. But this leads us onto the second stage of Foster's argument — that epiphenomenalism would rule out any possibility of language being used to refer to inner states. If this can be made out it would do two things — firstly, strengthen the argument that epiphenomenalism could not be asserted (purported assertions would not just fail to express knowledge, the crucial terms in them would fail to refer at all); and secondly, it would undermine the idea that epiphenomenalism could allow for even a private incommunicable knowledge of one's own inner states (assuming this is meant to be linguistically articulated knowledge). The argument was that words cannot be taken to refer to conscious states if there is no causal connection between the use of those words and the states they are purportedly about. Chalmers considers an argument of this sort. As with the previous argument about knowledge, he supposes that it takes a general causal theory as a premise. He briefly argues against a purely causal theory of reference, in favour of a what seems to be a more Fregean account: 'In referring to an entity all that is required is that our concepts have *intensions* . . . that the entity might satisfy' (Chalmers, 1996, p. 200). We don't necessarily have to be causally related to the entity in question.

It might seem from this that whether or not the argument about epiphenomenalism goes through will have to wait on complex and controversial discussions in the philosophy of language. One way of sharpening the issue would be to press the more radical challenge I mentioned above, and ask whether we can think of language having any reference or meaning at all if its production is to be explained in purely physicalistic terms. As Quine was happy to argue, a consistently naturalistic approach to language will undermine the very notion of meaning, leaving us 'only a behaviouristic ersatz' of 'intuitive semantics' (Quine, 1960, p. 66). Quine's

[12] I don't want to suggest that memory is a purely causal concept. The sense in which my memories are *about* past events is not fully captured by saying that a memory is a present state caused by past events. But it seems clear that there is at least a causal element in memory, and that is enough for my argument.

behaviourism means that he sees no need to add anything mental to a purely physicalistic account of language production. The epiphenomenalist will of course insist that mental items still enter the picture as possible objects of reference, but they cannot, *ex hypothesi*, make any difference to the account of language *production*, which will remain an exclusively physical one. If Quine is right that a consistently physicalist account of language undermines the notion of meaning, this should be taken as a *reductio*, since his argument cannot very well work unless we take the language he is using to have meaning. I think an argument along these lines would present a severe challenge to the epiphenomenalist, and Chalmers gives no indication of how it might be met. But I want to concentrate on the less radical argument that epiphenomenalism would prevent reference to conscious states, even if it didn't undermine the idea of reference or meaning generally.

Does this argument need to depend on a causal theory of reference? I don't think so. All it needs is some recognition that reference is not a wholly inscrutable abstract relation, but must involve a sensitivity to the ways in which the entities we refer to can impinge upon us.[13] In order for me to be credited with the ability to refer to xs, I must show that I can differentiate between situations in which I am justified in asserting the presence of xs and those in which I am not. I need to have some recognitional capacity. Now, if the occurrence in me of conscious states has no direct causal relation to my utterances of them, it seems hard to see how I could be thought to have any such capacity with respect to them. My utterances would, on the epiphenomenalist view, be disconnected so far from my inner states that they couldn't count as referring to them; they simply wouldn't have the necessary sensitivity to their presence or absence. (Of course, those utterances are sensitive to the brain states that are causally responsible for the production of the conscious states, but for epiphenomenalism this spinning off of consciousness is a spare-time activity in which these brain states indulge; what is that to the utterances? All they 'know' is that the brain state occurs, that it produces consciousness as a side-effect is irrelevant to the utterance.)

This is enough to ruin epiphenomenalism, since it means that it cannot be publicly asserted. It might still seem that each of us could give a private reference to our own conscious states, additional to the purely behaviourist/functionalist meaning our utterances may have in public. Foster simply comments: 'I cannot see how such private interpretations could have any bearing on the objective meaning of the terms, as employed in speech and writing, if, with respect to this employment, they are causally idle.' (Foster, 1991, p. 191) Such pure private bestowals of meaning would be irrelevant to the public use of language, even if they were possible; and since I have more sympathy for Wittgenstein's Private Language Argument than Foster does, I would add that they would not be possible in any case, for familiar Wittgensteinean reasons. I would conclude that Chalmers, at any rate, has not succeeded in disposing of the 'paradox of phenomenal judgment'; and that Foster's argument seems to go through. In which case, if epiphenomenalism can be asserted, it must be false. Foster, as noted above, doesn't think that this strictly refutes epiphenomenalism, although it does mean that it cannot be taken seriously and must be dismissed. But if we simply add, as a second premise, the contingent but undeniable fact that epiphenomenalism can be

[13] I should say that I am concerned here purely with reference to empirical entities (states, events . . .), and not with the problem of reference to 'abstract entities'.

asserted, can be discussed and talked about, then it follows straightforwardly that it is false. To repeat the moral of Section I, if this is true of Classic, dualistic epiphenomenalism, it is true of 'non-reductive' materialism also.

IV

If this is correct, where do we go from here? I noted above that epiphenomenalism is a conjunction of two propositions:–

(1) Consciousness is real and irreducible.
(2) Consciousness has no causal efficacy.

If epiphenomenalism has to be rejected, then at least one of these claims must be false. Eliminative or reductive materialists will deny (1). I shall not discuss this option here, except to note that if the more radical argument mentioned above — that a mechanistic or physicalist account of language undermines the very notion of language as bearing meanings — is correct, then it will follow that if such views are expressible at all, they must be false. It is occasionally argued in favour of reductive materialism that it is closer to common sense than the non-reductive version, in that it does allow for mental states to be causally effective. But this cannot be more than a bad joke. On the reductionist view, there is ultimately nothing more to a mental state than what is included in the physical description of it — so mental states are causally effective only because there isn't, ultimately, anything distinctively *mental* about them at all.[14]

Assuming that radical materialism is false, and that (1) therefore stands, we can either reject (2) or accept that we are stuck with an intractable antinomy. Is there any reason why we shouldn't reject (2)? Earlier on I argued that it followed from

(3) The physical world is a causally closed system.

If this is right, then to reject (2) would mean rejecting (3) as well, and it is often rather vaguely supposed that it would be 'unscientific' to do that.[15] I noted above Jackson's suggestion that to deny (3) would be equivalent to believing in fairies. Chalmers claims that such denial 'requires a hefty bet on the future of physics, one that does not currently seem at all promising; physical events seem inexorably to be explained in terms of other physical events' (Chalmers, 1996, p. 156). Is this right? (3) is not necessarily determinism, since it allows for the possibility that there could be physical events which were not explicable at all. But its adherents do generally seem to be in thrall to the picture of the world as a Great Machine. Quantum Mechanics is obviously an embarrassment, but it is supposed that its indeterminism can be kept neatly fenced off and prevented from contaminating the macroscopic realm. Thus Chalmers:

---

[14] Reductionists do characteristically insist that they affirm the reality of mental events. But to assert the identity of a mental and a neural event in a reductionist (as opposed to merely a dual-aspect) sense must involve the assertion that the physical description of the event is privileged; it is that rather than the mental description that tells us what the event *really* is. This is only compatible with preserving the reality of the mental if one starts by adopting a bleached-out 'topic-neutral' account of mentality — and this, at least as far as conscious mental states are concerned, is quite hopeless.

[15] Of course, if (2) doesn't follow from (3) we could solve the problem by rejecting (2) (and keeping (3)). But I have already argued — though, I agree, not entirely conclusively or comprehensively — against this option. What follows is premised on the assumption that one cannot deny (2) without denying (3).

'The physical world is more or less causally closed, in that for any given physical event, it seems there is a physical explanation (modulo a small amount of quantum indeterminacy)' (Chalmers, 1996, p. 150).

This view depends on a kind of scientific realism, according to which the laws of physics are to be taken literally as descriptions of the way the world ultimately works. But this way of thinking about physics is, to say the least, contentious. Even if we stick to classical physics, Newton's Laws do not simply describe the way the world works. They describe, in the first place, the workings of highly idealized and abstract models of aspects of the world. Relating these models to reality is a messy, ad hoc business, one that requires the laws themselves to be hedged around with *ceteris paribus* clauses. As Nancy Cartwright puts it: 'fundamental laws do not govern objects in reality; they govern only objects in models . . . Nature tends to a wild profusion which our thinking does not wholly confine . . . Things are made to look the same only when we fail to examine them too closely' (Cartwright, 1983, pp. 18–19). John Dupre, similarly, has argued that if we look at scientific practice we see nothing like the philosophers' tidy Unity-of-Science picture, with everything reducing down to deterministic laws of physics. We see lots of mutually irreducible types of explanation, which are generally probabilistic rather than deterministic. No one has yet come up with a good a priori argument for determinism and, according to Dupre, 'it seems almost entirely, or perhaps entirely, devoid of empirical support' (Dupre, 1993, p. 184).

Of course, such complex matters cannot be settled by a few citations. But I quote these remarks from distinguished contemporary philosophers of science in order to suggest that the picture of the Universe as a giant machine in which everything is explained by basic physical laws (modulo that little bit of awkward quantum stuff) is, at least, far from obviously correct, and far from being the only picture of the Universe which is scientifically respectable. The picture of the Great World-Machine is, in fact, very much the sort of picture which Wittgenstein warned could hold us captive (Wittgenstein, 1958, #115). We have a vivid, over-simplified image before our minds, and it becomes impossible to think beyond it. But this image is hardly rendered compulsory by anything in the sciences themselves, and, conjoined with the fact that consciousness is irreducibly real, it leads to the untenable position of epiphenomenalism. It is, accordingly, a picture which we should regard with deep scepticism.

We should not, therefore, be intimidated by vague claims that belief in mental causation is somehow 'unscientific'. However, it is still natural to ask: How do conscious states have causal effects on physical states? And Chalmers thinks that, by pushing this question hard enough, he can show that interactionist dualism doesn't really escape the problems with which it charges epiphenomenalism.

> Imagine (with Eccles) that 'psychons' in the non-physical mind push around physical processes in the brain, and that psychons are the seat of experience. We can tell a story about the causal relations between psychons and physical processes . . . without ever invoking the fact that psychons have phenomenal properties. (Chalmers, 1996, p. 158)

Chalmers does consider the natural response of an interactionist, that 'psychons' or whatever, are essentially or entirely phenomenal in their nature, so nothing would be left if we abstracted them from their phenomenal properties. To which he replies that 'even so . . . their phenomenal properties are irrelevant to the explanation of behaviour; it is only their relational properties that matter in the story about causal

dynamics' (*ibid*.). Hence the interactionist is still really in the same boat as the epiphenomenalist.

To this the interactionist could reply that it is irrelevant to the causal powers of brain events that they have conscious aspects, or cause conscious events, since it is in virtue of their other aspects that they are able to have physical effects. But in the case of 'psychons' they don't have any non-conscious aspects, so it must be in virtue of the conscious aspects that they can cause happenings in the brain. There can't be 'bare' relational properties, after all; a thing (event etc.) is able to enter into relations with other things (events) because of the intrinsic properties it has, and, *ex hypothesi*, all the intrinsic properties of psychons are phenomenal ones. However, Chalmers' point is a little more awkward for the interactionist than this response allows. For, after all, I do not usually make conscious decisions to bring about some pattern of neuronal firings in my brain — my decisions are to do things like having a drink. I am not aware of affecting my neurons in any way when I decide to take a drink, stretch out my hand and lift the glass. In which case, it might seem that we would have to assert that my decisions do have, somehow attached to them, properties unknown to me, and in virtue of which they have the causal effects on the brain that they do. But since I have no idea of these properties it would seem that they are acting in relative independence of my conscious states, and could theoretically be separated off from them.

This debate is not obviously conclusive either way, but I do think Chalmers is on to something. The underlying point of his argument is related to Nagel's claim that dualism doesn't solve the problem of 'what it's like', since it remains mysterious how even an immaterial substance could be me (see Nagel, 1986, p. 29). The intuition at the back of such arguments is that dualism, at least in its conventional, Cartesian forms, objectifies the subjective. Subjectivity, consciousness, what-it-is-likeness, cannot be fitted into the physical world. So it is itself reified, turned into a thing — albeit a thing which lacks the properties normally associated with things — and added to the standard physicalist ontology as a bolted-on extra. Even if arguments like Chalmers' don't show that this is impossible, the whole notion does seem distinctly unappealing, even to those of us without any sympathy for materialism or the 'scientific world-view'. Indeed, the defect of conventional dualism is that it is still too close to materialism; rather than making a radical break with it, dualism merely adds another element to it, and one that is itself characterized as a 'substance' or 'thing' — that is, in terms taken from our language for the objective physical world.

The defect of the sort of dualism that posits 'psychons' interacting at definite moments and points with the brain, is a sort of category mistake. Having correctly noted that conscious states (e.g. decisions) do cause physical events (e.g. the reaching out of an arm to a glass) the dualist tries to explain this at the level of brain functions. Again, this shows how much dualism is in thrall to the 'scientific world-view', sharing, in this case, its taste for bottom-up explanations. Since the brain states cause the overt bodily movements, dualism sees its task as one of trying to find an even deeper explanation — one that will suggest a causal mechanism to explain the brain activity. But this attempt to find a deeper mechanism is misguided, and does, as we have seen, leave the dualist somewhat vulnerable to criticisms such as Chalmers'.

In fact, if we take care to present explanations on appropriate levels, psychophysical causation ceases to be such a mystery. All that is required is, firstly, that a piece of behaviour is explicable in ordinary intentional terms, by reference to beliefs,

desires, values, etc. Secondly, that if we try to trace the brain activity that sends out the relevant signals to our limbs etc. back to source, we do not find an adequate deterministic explanation for it in physical terms. So if we stick to a purely physiological level of description, we can find no explanation for why *this* chain of neural activity occurred at this moment and led to the movement of this limb. But if we look for an explanation on the intentional level of why the subject moved her limb, this is forthcoming. In which case we can say that the neural activity occurred because the subject decided to reach out and lift up the glass.

That mental causation should be thought of in this way was suggested back in the 1960s by Charles Taylor, who wrote:

> may it not be that the only regularities discoverable will be those which emerge when we characterise those [neural] events as *embodiments* of the corresponding thoughts and feelings? In other words, no regularities may come to light as long as we characterise the embodiments in terms of physiological properties, but only when we see them as vehicles of the thoughts and feelings concerned. (Taylor, 1970, p. 239)

This does not require the postulation of 'psychons' intervening at a particular point in the causal process at the level of neuronal firings. E.J. Lowe, who has worked out a theory along these lines in some detail, has suggested that if we trace the physiological causal chains back from an overt limb movement, say, into the brain, we are likely to find that 'the chains will begin to display a tree-structure . . . Moreover, trees emanating from different peripheral events . . . will soon become inextricably intertwined' (Lowe, 1996, p. 64). Specific causal chains leading to overt bodily movements would, on this account, emerge unpredictably from a 'maze' of 'background' neural activity. Their emergence would be explicable neither in purely physiological terms, nor in terms of the intervention at specific points of 'psychons'; but the overt activity which they serve to bring about would be explicable in intentional terms.

This model of mentalistic top-down causation does require 'taking a bet' that a fully deterministic account of the workings of the brain will not be produced. But it is hard to see how this bet could be lost; the brain is, after all, a system of such fantastic complexity that there is no serious prospect of ever demonstrating that it is a fully deterministic system taken as a whole. The problem isn't even a 'merely' practical one; at this level of complexity it seems that Chaos Theoretical considerations become relevant. The entanglement of Lowe's causal chains as one traces them back would soon ramify beyond even the theoretical possibility of prediction. Of course, from the lack of predictability, even from a 'Chaotic' in-principle-lack-of-predictability, we cannot infer that a system is not deterministic; but, since there are no good a priori reasons for supposing universal determinism to be true, then, in the absence of predictability we would have no good empirical reason for supposing it.

One final objection to the model I have proposed is that it fails to explain exactly *how* our conscious decisions impact on the brain. Actually, this seems to me to be a virtue. The error of the 'psychon' account was that it tried to explain mental causation on the wrong level. My account puts our intentional explanations to use in the place where we do actually use them; to explain our (more or less) intelligent behaviour, not our neuronal firings. I can see no a priori reason to suppose that there has to be a complete explanation for our neural processes at their own level; indeed, the main arguments of this paper were meant to show that there are good a priori reasons for not

expecting any complete explanation on this level. Those who remain dissatisfied seem to me to illustrate Wittgenstein's dictum: 'The difficulty . . . is not that of finding the solution, but rather that of recognising as the solution something that looks as if it were only a preliminary to it . . . The difficulty here is: to stop.' (Wittgenstein, 1981, #314)

Of course there are plenty of large issues which need to be discussed further. I don't take myself to have *proved* anything in this final section. But I do hope I have done something to show that to reject mental causation and insist on the causal closure of the physical world — despite the deep problems and incoherences into which that leads us — is unreasonable and based on certain philosophical prejudices. If a really watertight case could be made for the causal closure of the physical, then, assuming the correctness of my case against epiphenomenalism, and the falsehood of radical (reductive/eliminative) materialism, we would be faced with a peculiarly harsh and troubling antinomy. But since there is no such watertight (or even half-way plausible) case for causal closure, I would conclude that a belief in mental causation along the lines suggested is the most reasonable option.

References

Cartwright, N. (1983), *How the Laws of Physics Lie* (Oxford: Clarendon Press).
Chalmers, D. (1996), *The Conscious Mind: In Search of a Fundamental Theory* (Oxford: Oxford University Press).
Davidson, D. (1980), 'Mental events', in D. Davidson, *Essays on Actions and Events* (Oxford: Clarendon Press).
Davidson, D. (1993), 'Thinking causes', in *Mental Causation*, ed. J. Heil and A.R. Mele (Oxford: Clarendon Press).
Dennett, D. (1991), *Consciousness Explained* (Harmondsworth: Penguin).
Dupre, J. (1993), *The Disorder of Things: Metaphysical Foundations of the Disunity of Science* (Cambridge, MA: Harvard University Press).
Foster, J. (1991), *The Immaterial Self; A Defence of the Cartesian Dualist Conception of the Mind* (London and New York: Routledge).
Hutto, D. (1999), 'A cause for concern: reasons, causes and explanations', *Philosophy and Phenomenological Research*, **59** (2), pp. 381–401.
Jackson, F. (1990/1982), 'Epiphenomenal qualia', in *Mind and Cognition*, ed. W. Lycan (Oxford: Blackwell). Reprinted from *Philosophical Quarterly*, **32**, pp. 127–36.
Kim, J. (1989), 'The myth of non-reductive materialism', *Proceedings and Addresses of the American Philosophical Association*, **63** (3), pp. 31–47.
Kripke, S. (1980), *Naming and Necessity* (Oxford: Blackwell).
Lowe, E.J. (1996), *Subjects of Experience* (Cambridge: Cambridge University Press).
McLaughlin, B. (1993), 'On Davidson's response to the charge of epiphenomenalism', in *Mental Causation*, ed. J. Heil and A.R. Mele (Oxford: Clarendon Press).
Malcolm, N. (1968), 'The conceivability of mechanism', *Philosophical Review*, **78** (1), pp. 45–72.
Nagel, T. (1986), *The View From Nowhere* (Oxford: Oxford University Press).
Quine, W.V. (1960), *Word and Object* (Cambridge, MA: MIT Press).
Taylor, C. (1970), 'Mind–body identity — a side issue?', in *The Mind/Brain Identity Theory*, ed. C.V. Borst (London: Macmillan). (Originally, *Philosophical Review* (1967).)
Wittgenstein, L. (1958), *Philosophical Investigations*, trans. G.E.M. Anscombe (Oxford: Blackwell, 2nd edn.).
Wittgenstein, L. (1981), *Zettel*, trans. G.E.M. Anscombe (Oxford: Blackwell, 2nd edn.).

Paper received January 1999