Quantum Curiosities of Psychophysics, by **Jeremy Butterfield**

Abstract: I survey some of the connections between the metaphysics of the relation between mind and matter, and quantum theory's measurement problem. After discussing the metaphysics, especially the correct formulation of physicalism, I argue that two state-reduction approaches to quantum theory's measurement problem hold some surprises for philosophers' discussions of physicalism. Though both approaches are compatible with physicalism, they involve a very different conception of the physical, and of how the physical underpins the mental, from what most philosophers expect. And one approach exemplifies a a problem in the definition of physicalism which the metaphysical literature has discussed only in the abstract. A version of the paper has appeared in *Consciousness and Human Identity*, ed. John Cornwell, OUP 1998.

1. Introduction

My subject is quantum theory and the mind. These are very disparate topics; so it can seem foolhardy to link them. But I believe there are some genuine connections between them. More specifically, I believe there are connections between two philosophical aspects of them: between the problems about interpreting quantum theory, especially the so-called 'measurement problem', and the metaphysics of the relation between mind and matter. In this paper, I will survey some of these connections. I will emphasise the metaphysics, especially in the first half of the paper (Sections 2 and 3); turning only in the second half (Sections 4 and 5) to how quantum theory's measurement problem bears on the metaphysics. (Two complementary papers emphasise the quantum theoretic issues; and discuss 'no-collapse' solutions to the measurement problem, which I set aside here: 1995, 1996).

I will have two main conclusions. They are both about how some ways of solving the measurement problem, i.e. some interpretations of quantum theory, bear on the doctrine of physicalism. This is the view, very roughly speaking, that everything is physical: less roughly, that all facts, and *a fortiori* all mental facts, are actually physical facts.

My first conclusion is that although some interpretations of quantum theory, which I will present, are compatible with physicalism (according to various exact definitions), they involve a very different conception of the physical, and of how the physical underpins the mental, from what most people expect—including philosophers and neuroscientists participating in the debate about whether the mind is physical. In short: the interpretation of quantum theory holds some surprises for this debate. Hence my title!

My second conclusion is more specific. It is that one interpretation of quantum theory gives a real-life example of a problem, which metaphysical discussions of physicalism have seen only in the abstract. It is a problem about how exactly to define physicalism. Namely: one attractive and much-discussed strategy for making physicalism precise threatens to be too weak. That is, the precise definition can be satisfied, even while intuitively physicalism is false. So my point will be that one interpretation of quantum theory illustrates this problem.

To establish these conclusions, I will first need to summarize some aspects of the contemporary debate about the mind-matter relation, leading up to the much-discussed definition of physicalism. (To anticipate for a moment, using the jargon of philosophy: it defines physicalism as a contingent supervenience thesis.) This summary is in Sections 2 and 3.

Section 2 clears the ground. It has three subsections. In these, I argue that: there is a *prima facie* distinction between mind and matter (Section 2.A); that some mental states are literally in our heads (Section 2.B); and that even if these mental states are identical with states of our brain, no dubious 'reductionism' follows (Section 2.C).

Section 3 formulates physicalism. It has four subsections. In Section 3.A, I distinguish physicalism from the logically weaker doctrine of materialism; (the doctrine, very roughly, that everything is completely describable by the natural sciences). In Section 3.B, I distinguish two senses in which it could be true that 'all facts are physical facts'. The first, I call 'reduction'; the second, 'supervenience'. Reduction is logically stronger: that is, it implies supervenience. In Section 3.C, I spell out the details of these senses—and meet an obstacle. Namely: if we consider only actual facts, physicalism in these two senses is in grave danger of being trivially true, as a matter of mere accident. In Section 3.D, I describe how to overcome this obstacle: we need to consider not only the actual facts, but also possible facts—how the history of the universe might have gone. But exactly which possible facts? One much-discussed answer is: ‘the possible facts that accord with the actual laws of nature’. It is this answer that faces the problem mentioned above. It makes the formulation of physicalism too weak: physicalism comes out true in cases where intuitively the mental is *not* physical, but is just rigidly connected to the physical by laws of nature.

That will conclude my general discussion of physicalism. Section 4 turns to quantum theory. First, in Section 4.A, I warn against a naive picture of physical reality as composed of particles like billiard balls moving in a void. This picture has led to bad arguments about physicalism. In particular, I rebut a bad traditional argument *for* physicalism, which is still endorsed by some authors.

In Sections 4.B and 4.C, I describe quantum theory’s measurement problem. Roughly speaking, it is this: quantum theory's laws about how the states of objects change over time seem committed to the prediction that macroscopic objects often have no definite positions—nor definite values for other familiar physical quantities like momentum or energy. But this seems manifestly false: tables and chairs surely have definite positions etc. As it is sometimes put: the macrorealm is definite. Or at least, we experience the macrorealm as being definite. So, if quantum theory is to account for our experience, it must either secure such definiteness, or at least explain the appearance of it.

Section 5 discusses one approach to solving this problem. This approach postulates new physical laws which avoid the false prediction. (I set aside another approach, in which the ordinary laws of quantum theory are retained, but the false prediction is avoided by postulating ‘extra values’ for some physical quantities. Suffice it to say that, as in Section 5, some versions of that approach are compatible with physicalism, but others are not: cf. my 1995, esp. pp. 145f.).

I will consider three versions of this approach. The first two versions (Section 5.A) are compatible with physicalism. But they yield my first conclusion. That is: on these proposals, the *way* in which the mental is fixed by the physical would be very different from what most physicalists, unversed in the controversies about quantum theory, would expect. The third version (Section 5.B) yields my second conclusion. Here is an interpretation of quantum theory which illustrates the problem at the end of Section 3.D: an interpretation that violates the spirit of physicalism, if not its letter.

2. Mind and Matter

In this Section and the next, I will summarize some aspects of the contemporary debate about the mind-matter relation. I have chosen these aspects with a view to complementing other papers in this volume, especially those by Boden, Lipton and Searle; whom this Section will try to locate in the debate.

We have countless beliefs about the material (empirical) world in space and time; including our bodies. And we have countless beliefs about the mental world: that is, the minds and experiences of ourselves and others, including perhaps animals. So here are two subject-matters for our beliefs, each very large and heterogeneous. (And they are also subject-matters for our other attitudes: desire, question, etc.) The question arises: What is the relation between these subject-matters? That is our problem, the so-called ‘mind-body problem’; though because ‘body’ has the specific connotation ‘animal organism’, rather than ‘material object’, it would clearly be better to call it the ‘mind-matter problem’. As stated, the problem is vague: we have no precise notions of subject-matters, and of relations between them. Once we clarify those notions, the problem will no doubt break up into different problems, some scientific and some philosophical.

In the first place, it is not easy to state exactly what these two subject-matters are. And accordingly, many authors take a sceptical view of the whole debate. So I first address this view.

2.A: Denying the Mind-Matter Distinction

These authors doubt that a precise distinction between the subject-matters can be made; or they doubt that once it is made, it is scientifically and/or philosophically significant—that it in some way ‘carves nature at the joints’ (For example, Midgley (this volume) argues that materialism and idealism are both half-truths, to be transcended.)

This scepticism is often supported by two related lines of argument. First, one can spell out the historical process by which this distinction entered our philosophical culture. The orthodox idea is of course that it entered through the mechanical philosophy of Galileo, Hobbes, Descartes and their successors, who shared a common vision of a mathematical mechanics underlying all material facts, though perhaps not that apparently very disparate arena—the mind. In spelling out this historical process, one discovers, unsurprisingly, that there was less of a common vision than this orthodoxy suggests; and one gets a vivid sense that our contemporary mind-matter distinction is historically moulded, and is not an intellectual necessity (e.g. Baker & Morris 1993). Both points tend to undermine one's conviction in the scientific or philosophical importance of that distinction.

Second, one can argue that the development of science since 1700 has not vindicated the alleged common vision. There is an uncontentious point here: the concept of a mathematical physics, as the 'basic science', has changed drastically, above all by giving up the primacy of mechanical concepts.[1]

But this second line of argument also makes a contentious claim: that physics, even as transformed since 1700, is in no sense 'basic'; that on the contrary, modern science is in fact very disunited, with the separate sciences pursuing their own goals, and developing their own theories, entirely regardless of physics (e.g. Dupre 1993). Such authors often go on to argue that this disunity seems likely to continue—it is not a temporary predicament. So they propose to drop the binary mind-matter distinction, as an artefact of a certain philosophical legacy; and to replace it by a picture of many subject-matters studied by the many separate sciences. These subject-matters are no doubt connected in various ways. For example, by causation, i.e. an event or state of affairs in one subject-matter causing an event or state in another; and by laws of nature, i.e. there being a law of nature relating events or states of affairs in two or more subject-matters . But, say these authors, these subject-matters are not all somehow dependent on some basic subject-matter (that of physics), nor somehow arranged in a heirarchy or spectrum of 'basicness'.

My own position (of course, shared with many authors) is that the mind-matter distinction *can* be made good; and furthermore, once it is made good, this last claim is wrong. That is, physics is indeed 'basic', and modern science is well-unified. At first sight, this position is likely to seem 'reductionist' or 'eliminativist'. But as I develop this position, in this Section and the next, we will see that in fact it is not (at least in the most usual senses of these notoriously vague words!).

Given the distinction, there are, as I said above, various problems to address, some scientific and some philosophical. The most exciting scientific problem is perhaps (as suggested by Searle (this volume)) how mental states, especially conscious states, are biologically produced. I of course agree with Searle that for this problem, the philosopher's role is primarily to help clear away some confusions. Thus also Locke, who said in the 'Epistle to the Reader', in his *Essay* that he was content to be 'an underlabourer ...clearing ground a little, and removing some of the rubbish that lies in the way of knowledge' (1972, p.xxxv)

As to the philosophical problems, I shall concentrate on what I take to be the main metaphysical issue: in what sense, if any, do neural states 'underly' mental states? Lipton (this volume, Section 1) proposes that the two main candidates for this relation of 'underlying' are: causation and identity (cf. also Boden (this volume) Section II). In Section 3, I will advocate the second candidate, identity. (Lipton's Section 2 summarizes some difficulties confronting this view, and recommends scientists to be content with causation. But as will emerge, I agree with most of what Lipton says.)

But, as for the mind-matter relation in general, I should first address a sceptical view: that in no sense do neural states underly mental states. (Doing so will occupy most of this Section.)

2.B: Denying that Neural States Underly Mental States

This sceptical view is based on the point that mental states, as we ordinarily conceive them, involve the world: in the jargon, they are 'wide'. This width is shown by some standard examples from recent philosophy of mind. Thus authors such as Burge, Kripke and Putnam argue that you could not believe that London is pretty, that water is wet, that arthiritis is painful, just in virtue of your brain state. For you can only be credited with the concepts occurring in these beliefs, concepts like London, water and arthiritis, if your environment has certain features, and if you are appropriately related to those features. In other words, a Doppelganger of you, living on another planet, would not have these concepts, no matter how much their neural states matched your neural states, unless certain features of their environment, and their relation to those features, similarly matched yours. Furthermore, these features of the environment need not be physical features, nor features that can uncontentiously be spelt out in physical (or more generally, natural scientific) terms. They can be social features, e.g. the meaning of 'arthiritis' in a language. And the appropriate relation need not be a physical relation; nor need it be a relation that can uncontentiously be spelt out in physical, or natural scientific, terms. It might be e.g. membership of a linguistic community. (Cf. Burge 1979, Kripke 1979; Putnam 1975).

I take it that these points are established by these authors' work. Thereafter, matters become difficult: it is very hard to say for a given concept, exactly what is required of the environment, and of your relation to it, if you are to have the concept. For example: to have the concept of London, must you have had causal contact (however exactly that is defined!) with London? But of course, these difficulties do not need to be solved in order to attack the idea that neural states underly mental states. For that, the established points are enough.

Indeed, these very difficulties enable one to generalize the attack. Thus, it can seem plausible that however exactly you define 'mental' and 'physical', mental and physical facts are intertwined in the world in so myriadly complex a way, that no limited class of physical facts 'underlies' (in any sense) all the mental facts—and that this remains so, even when 'physical' has the very general meaning 'natural scientific', and even when one goes beyond the body of the individual person or animal. (In going beyond the individual, one immediately meets the social dimension of mind; here, talk of mental, and indeed social and cultural, 'facts', is more common than talk of 'states': hence this paragraph's change of terminology.)

In other words, it seems that not only do the mental facts about a person, or animal, 'outstrip' the facts about its nervous system: they outstrip all the physical, chemical and biological facts about it—and even those of other persons or animals. To adapt Midgley's example of treaties (this volume, p.?? (p.21 of MS)): it seems that no collection of natural scientific states of a nation's citizens underly the nation's agreeing to a peace treaty.

This view is of course close to the sceptical view I discussed in Subsection A, that either the mind-matter distinction cannot be made, or once it is made, it is without significance. The difference is that this view is less radical. It allows that the mind-matter distinction may be made (perhaps in various ways), and have various kinds of significance. But it denies one much–discussed kind of significance—that the physical (even in the sense of 'natural scientific') 'underlies' the mental.

My reply to this view (again, shared with many authors) consists of two parts. Though I endorse both parts, they are independent: one can endorse each part without the other. The first part is directed at the generalized attack; and is, I admit, wholly programmatic. In effect, it consists of two broad strategies for finding the physical facts underlying some mental (or any other, e.g. social) fact, such as acceptance of a peace treaty; as follows.

(1): To find these physical facts, we may need to look far more widely across the environment, both in time and space, than is initially suggested by a sentence reporting the fact. Thus the physical facts underlying the social fact 'Napoleonic France accepted the treaty of the Congress of Vienna' will surely not be confined to France and Vienna, even in the year 1815. As to time, the physical facts underlying France (or any nation) accepting a treaty may well include facts earlier even than the lifetimes of the relevant politicians, or other citizens—who knows how far back in time one has to go, in order for the physical facts to fix our concept of a treaty? As to space, a similar point can be made.[2]

(2): To find these physical facts, break the problem up. That is: think of some sequence of kinds of fact, which has some kind of physical facts as the first member of the sequence, and for which it is plausible to claim that each kind of fact underlies the kind that next occurs in the sequence—and then argue *seriatim* for these claims. As thus stated, this strategy suggests the traditional reductionist heirarchy, with physics reducing chemistry, chemistry reducing biology, biology reducing psychology, and psychology reducing the social sciences (Oppenheim & Putnam 1958; Monod, 1972). But this strategy is by no means committed to this heirarchy, regarded by many as a bugbear: for at least three reasons (and others will emerge below).

First, 'underly' need not mean 'reduce' (at least in the sense of 'reduce' that you consider a bugbear!). Second, the strategy allows different sequences of kinds of fact (maybe criss-crossing), rather than just a single line going 'up' from the physical. Third, the strategy allows very different kinds of fact (as well as sequences) than this traditional list.

Here, it is worth mentioning a well-worked out example of the strategy that uses such different kinds of fact; for although the example is well-known to analytic philosophers, it is almost unknown outside philosophy. Namely, David Lewis' strategy for establishing what he calls 'Humean supervenience' (Lewis 1986, pp. ix-xiv). Roughly speaking, his plan of battle is as follows. He argues that the physical facts underlie (i) facts about laws of nature, and (ii) facts about counterfactual conditionals (though neither of (i) and (ii) underlies the other). Facts about counterfactual conditionals then underlie (iii) facts about causation. Then (i) and (iii) underlie (iv) largely individualistic facts about the mind (via a version of functionalism). Then (iv) underlies linguistic facts; (here the argument includes an analysis of linguistic convention in terms of mental facts about e.g. common knowledge of how symbols would be interpreted). The arguments for these claims are impressively detailed—so much so that one can indeed believe that physical facts underly a nation's accepting a peace treaty: I recommend them to sceptics like Midgley. But we must forego the details of these arguments. For now, it is enough that we can already see how different are these kinds of fact from the traditional list above. (And the further details of Lewis' battle-plan reveal further differences.) To sum up: this strategy can be very imaginative!

I turn to the second part of my reply to the sceptical view (again: shared with many others). It is directed at the arguments and examples of Burge et al., and is more specific. In contrast to the 'look more widely' of (1) above, it says in effect 'look more narrowly'. That is: I concede that mental states as ordinarily conceived are wide (as the examples suggest), but still claim that a person (or other animal) has *some* mental states just in virtue of their neural states. Such mental states—ones that would be shared with an organism with exactly matching neural states—are called 'narrow'. The usual candidates are states of sensory experience: something like seeing yellow in the top left of one's visual field. Agreed, such states are hardly mentioned in everyday life. But there is an obvious pragmatic reason for this. We are primarily interested in describing, deciding, predicting and explaining our own and other people's actions: and here 'action' means not just the 'narrow' idea of body-movements, but these movements' causal consequences in the world. With actions thus widely conceived, describing, deciding, predicting and explaining them naturally mentions wide mental states. Thus the advocate of narrow mental states can, and no doubt should, allow that such states are hard to describe in everyday language, which tends to use a 'wide' taxonomy of states. Maybe even 'yellow', as a word of everyday language, expresses a concept whose possession requires something of one's environment: (no doubt, nothing so strong as having seen a lemon or banana—but maybe one has to have seen a yellow object).

But despite these difficulties of description, I think that there are such narrow mental states, and that states of sensory experience are examples; so that some narrow states form a constant accompaniment to the wide states that we usually ascribe to ourselves and other people.[3]

2.C: Physicalism with a Human Face

Given such narrow mental states, the way is clear to discussing the main metaphysical issue introduced above: exactly what is this relation of 'underlying', between neural states and narrow mental states?

In Section 3, I will address this question: namely by formulating physicalism as a thesis of supervenience. Very roughly speaking, physicalism will say that physics is indeed the 'basic' science. A bit more exactly: physicalism says that as a matter of contingent fact, the physical facts about the cosmos completely fix all the facts. This will make the relation of 'underlying' be identity, i.e. neural states are narrow mental states. (I will also suggest that given their other views, Searle and Lipton can and should agree.)

But before embarking on the details of formulating physicalism, I want to emphasise that my physicalism does not imply various dubious doctrines. First, it will turn out that my physicalism is not 'reductionist' or 'eliminativist', in most senses of these words. We can already state one reason why not. It lies in the distinction between metaphysics, on the one hand, and epistemology and explanation on the other. Thus my metaphysical thesis of physicalism, and the thesis that neural states are (narrow) mental states, implies very little about the epistemology of mental states (e.g. how can we know about them, and what are our best methods for getting such knowledge?), and very little about the character of explanations of mental states, and processes (i.e. sequences of states). In particular, it does not imply that physical explanations of these states or processes are somehow 'better'. A principal reason for this is the fact that, as Lipton rightly emphasises (this volume, Section 3), most explanations are contrastive; and once we recognize this, we can overcome the illusion that there is, or should be, a 'complete explanation' of something, such as a mental state. (As he notes, this point of view is also endorsed by Boden, Midgley and Watts.)

Similarly, I maintain that my physicalism and the identity of narrow mental states with neural states do not imply the doctrines that Rose identifies as the main errors of reductionism: reification, arbitrary agglomeration, etc.— and which he rightly castigates as not merely false, but as leading to grotesquely wrong public policies (this volume). I shall not tackle these errors *seriatim*. Suffice it to say here that I think all but two of these errors arise from the naive desire to explain myriadly different phenomena in a simple, unified way, in terms of their smallest parts—a unitary model of explanation that Lipton shows us to be a chimera. The two exceptions are the errors of improper quantification and belief in statistical 'normality'. These obviously arise from our scientific culture's over-valuing mathematical models: which is itself caused, at least in part, by the success of mathematics within physics. But for all its grotesquely wrong social consequences, this over-valuing of mathematics is, from the strict viewpoint of the philosophy of the mind-matter relation, a 'merely' historical fact. And right or wrong, setting a high value on mathematical models is a methodological claim, and no part of the metaphysical thesis of physicalism.

3. Physicalism

I turn to formulating physicalism. It says, roughly speaking, that all empirical subject-matters, such as the biological, the mental, the social, are literally a part of the subject-matter of physics. I shall make this precise as a so-called 'supervenience' thesis, where supervenience is a relation between subject-matters. Although several authors (including physicalists!) have recently criticized this approach to formulating physicalism, I believe it is viable—and that physicalism so defined is plausible. But unfortunately, I have space only to spell out the approach (in a version that owes much to the metaphysical system of David Lewis); not to address the criticisms.

3.A: The Material and the Physical

Before explaining the crucial relation of supervenience (Section 3.B et seq.), I need to make three preliminary remarks, which will guide the precise formulation of 'physical' and of 'supervenes'.

(1) A claim more modest than physicalism is often discussed: it is usually called 'materialism'. Again speaking roughly, materialism claims that all empirical subject-matters are part of the subject-matter of the natural sciences. So materialism does not require that they all be part of physics: one could be part of physics, another part of biology etc. Although this is weaker than physicalism, it is still quite radical! For, taking just the case of interest to us, the mind: this seems a very different subject-matter from those of the natural sciences. The two general differences most commonly cited in philosophy are that two concepts (i.e. properties) that seem important for describing minds and experiences seem entirely absent from all the natural scientific theories (in physics, chemistry, biology) that we use to describe the material world. Namely, the concepts of: (i) intentionality—the concept that a belief or perception or desire or some other mental state is about an object or state of affairs; and (ii) qualia—the concept that a conscious mental state, like a perception of an object outside the body, or a bodily sensation, has a 'raw feel', 'phenomenal quality', or 'quale'. Materialism (and *a fortiori,* physicalism) is committed to claiming that these differences are illusory: that a natural scientific account of intentionality and qualia is possible. To put the same claim in terms of subject-matter, instead of its linguistic description: the concepts (i.e. properties) that seem distinctively mental are in fact material -- though perhaps very complicated, or in some other way special.

(2) Materialism has been widespread in philosophy since the late nineteenth century, as a result of of the great success, since about 1800, of natural scientific theories in predicting and explaining the material world. One well-known example is the demise of vitalism within biology. Here are some examples from astronomy, chosen with a view to emphasising how it gradually emerged that the same laws govern processes on earth, and far away in space—a remarkable unity in the material world, which nowadays we tend to take for granted. In 1793, Herschel showed that double stars circled each other, confirming that Newton's law of gravitation applied outside the solar system. From 1860, Kirchoff and Bunsen applied their spectroscopes to sunlight, and thus paved the way to determining the chemical constitution of the stars (just 25 years after the positivist Comte gave this as an example of the sort of information that science could never attain!). And it emerged that all the elements detected in the stars were also found on Earth; though, to be sure, there were times when this seemed false, the best-known example being the 27-year gap between the discovery of helium in the Sun (1868: hence the name!), and on the Earth (1895).

Similarly, physicalism is widespread in present-day philosophy, as a result of the great success, since about 1900, of physical theories in predicting and explaining not only physical, but also chemical and biological processes. Here are two well-known examples, chosen with a view to emphasising how quantum theory can now claim to be the fundamental theory of matter—and how this claim has been hard-won. First: quantum theory's explanation of the homopolar chemical bond (achieved in 1927, just after the discovery of quantum theory, by Heitler & London); second, quantum theory's explanation of superfluidity and superconductivity (achieved by 1960, by the combined work of many).

Of course, I do not intend these historical points as persuasive arguments for the truth of materialism and physicalism. But I do think they indicate that precise formulations of materialism and physicalism should render them as contingent theses. For it is a thoroughly contingent fact that since 1800 there has been such supreme success in the natural scientific description of the world; and that since 1900 there has been such supreme success in physics. So surely our formulations should reflect this contingency of our intellectual culture.[4]

(3) This leads to the next remark. Today's natural scientific theories, and physical theories, are of course not wholly true. No doubt some of what they say is false. Furthermore, they are partial: they do not *completely* describe (truly or falsely) their own immediate subject-matter, let alone subject-matters like the mental and the social which materialism and physicalism claim to fall within their scope. So materialism and physicalism should be formulated so as to claim correctness and completeness only for some sort of corrected and extended versions of these theories. So the question arises: How exactly should we define these corrected and extended versions? The threat of triviality looms: we must not define them as whatever corrections and extensions are needed to make our formulations come out true. (As opponents of physicalism have often pointed out: e.g. Healey 1978, Crane and Mellor 1990).

This is an important, and hard, question, which I cannot answer exactly. It must suffice to make a few points. First, some authors sketch definitions in terms reminiscent of Peirce: they appeal to the long-run future of the natural sciences, or of physics—or rather, what this future would be, if circumstances were sufficiently propitious for enquiry (e.g. enough funding!). But I find such definitions too dependent on contingencies about what humans require for successful enquiry. I prefer more metaphysical definitions: specifically, those based on the claims that (i) we already have, from present-day theories, a good idea of what counts as a natural scientific, or physical, property—and (ii) we can readily enough make this idea precise. Given claims (i) and (ii), the corrections and extensions can be straightforwardly defined as the true complete theories of those properties.

So the issue turns on the claims. Obviously, the greater unity of physics, as against the various natural sciences taken together, makes claim (i) more plausible for physical properties, than natural scientific ones. (Since I favour physicalism, this difference of plausibility is no difficulty for me: materialism, being logically weaker than physicalism, is supported by whatever evidence supports physicalism.) I myself believe that our present idea of physical property has two main components; I admit that both components are vague (the second more so), so that establishing claim (ii) is hard—but for the present, they will have to do.

First, a physical property is a numerically measurable property of objects, whose value changes in time (if at all) according to a law which relates it to other such properties and their values; (typically, a differential equation or generalisation of such, like a stochastic differential equation).

Here and in what follows, ‘properties’ includes relations like ‘rotating faster than’ as well as monadic properties like ‘is electrically neutral’; and ‘objects’ is meant very generally, including what we would more naturally call events, processes or states of affairs.

Second, any such property is related to the physical properties we have already discovered, the number of which is astonishingly small, given the universal scope of physics. (Depending on how you count, there are somewhere between about a dozen, i.e. position, mass, electric charge, spin etc; and a hundred, i.e. including energy, momentum, entropy, temperature, conductivity etc.) This relation to the already-discovered properties is to be (or include) numerical relations between values, reported in equations. (This relation might be reduction or supervenience as defined below, but it need not be: physicalism should of course allow that a genuinely new physical property, i.e. not supervening on the already-discovered, might exist.)

3.B: Relations Between Subject-Matters: Reduction and Supervenience

So much for preliminary remarks. I turn to being more precise about relations between subject-matters. In accordance with remark (3) just above, I take a subject-matter to be a characteristic family (i.e. a set) of properties, defined on some set of objects. (Again: 'properties' includes relations; and ‘objects’ includes e.g. events, states of affairs.)

You might reasonably object that a subject-matter should include not only a set of objects, and a family of properties, representing a taxonomy (classification-system) for the objects; but also doctrines—general propositions about how the properties are related, e.g that all Fs are G, that no G is both an H and a K. (Such propositions might even be laws.) But no worries: I shall take the subject-matter to include each property, not just as a 'concept', but rather as an extension, i.e. as a set of instances. Since all such general propositions will be implicitly fixed by the extensions, e.g. the set of Fs being a subset of the set of Gs, they will in effect be included in the subject-matter.

Similarly, it will be convenient to speak below of two properties being identical with another, when they are co-extensive (i.e. have the same set of instances) in the given set of objects. At first sight, this usage seems to violate the standard view that intuitively distinct properties can be accidentally co-extensive; e.g. 'has a kidney' and 'has a heart'. But it is *just* a usage: I endorse this standard view. And furthermore, I will argue in Section 3.D that for the case of interest, i.e. physicalism, the set of objects needs to go beyond the actual world, i.e. to include objects that do not actually exist. As a result, this usage will not even *appear* to violate the standard view.

As mentioned in Section 2.A, one obvious way in which two subject-matters can be related is by causation and/or law, while neither is in any sense part of the other. Thus there might be causal relations between objects in the two sets; and there might be non-causal but nomic relations (from the Greek word for law, νομοσ). An example is classical electricity and magnetism; where each is thought of as, say, the family of possible values for the electric (or magnetic) field, defined on the set of spacetime points. There are certainly nomic relations between these values (given by Maxwell's equations); and maybe causal ones too, though the role of causation in physics, even classical physics, is controversial. But neither subject-matter is a part of the other; (though indeed, both are reduced to a single underlying subject-matter, the electromagnetic field). An example from everyday experience is France and England; where each is thought of as, say, the family of all empirical properties of the relevant set of citizens. Here there are plainly causal, as well as nomic, relations; but neither is part of the other.

For one subject-matter to be part of another, it is obviously necessary that the first subject-matter's set of objects be a subset of the second's. But that is not enough: in the example of electricity and magnetism, there is a single set of objects—the spacetime points—but two distinct subject-matters. Obviously, the first subject-matter's family of properties must also be somehow a part of the second's: every classification of objects made by the first must also be made by the second. The simplest way this can happen is if the first's family of properties is included in, i.e. is a subset of, the second's. (Since I take a subject-matter to include each of its properties' extensions, this inclusion will be enough to secure that the doctrines, i.e. general propositions, of the first subject-matter are included in those of the second.)

But typically, we are not so lucky: subject-matters are often given to us without one family being a subset of the other. A trivial example is squares and rectangles. Since squares are just a special case of rectangles, the subject-matter, squares, should surely be part of the subject-matter, rectangles. But the latter might not be given to us as including the property 'is a square', and all the various related properties, like 'is a diagonal of a square', that occur in the subject-matter, squares. After all, why should the subject-matter, rectangles, single out such special cases?

The remedy is not far to seek. We obviously need to use the idea of *compounding* properties to yield other properties, where the compounding operations can in general be iterated. Then we define one subject-matter to be a part of another iff:

(i) its set of objects is a subset of the other's; and

(ii) its properties either are among the other's properties (the simple, lucky case)—or are compounds of them.

(As I said above, clause (ii)'s notion of identity for properties and their compounds is straightforward: it is just coextension in the given set of objects.)

This definition focusses attention on compounding operations. About these, I should first make two brief, related points. They both arise from my taking a subject-matter to include each of its properties' extensions.

(1) It turns out that for the usual compounding operations, clause (ii) of the definition implies clause (i). (2) More important, we should no doubt require that if one subject-matter is a part of another, then its doctrines are in some corresponding sense a part of that other's, e.g. by being entailed by them. And you might reasonably worry that, even with subject-matters taken as including extensions, the above definition does not secure this requirement—might not compounding break out of the class of entailed propositions? But it turns out that for the usual compounding

operations, there is no problem: the definition implies that the doctrines of the first subject-matter are indeed entailed by those of the second.

So what are these compounding operations? The most simple are the so-called Boolean operations, represented by words like ‘and’, ‘not’ and ‘or’: giving, when iterated, all the Boolean compounds of the initially given properties. These operations will certainly suffice in simple cases like squares and rectangles, maybe even without iteration. For example, suppose the subject-matter, rectangles, is given as containing the properties, being a rectangle, and being a plane figure with four equal sides. The conjunction of these properties is the property, being a square. And once this link is made among properties, there is of course no problem about the deduction of the doctrines: you can deduce all doctrines about squares from all the doctrines about rectangles.

But there are more complex compounding operations. The two obvious examples from logic and philosophy are applying the quantifiers 'all' or 'some' to get a monadic property, e.g. 'bears relation R to everything/something', from the binary relation 'bears relation R to'; (or more generally, an n-adic relation from a (n+1)-adic relation). These two examples behave rather like the operations of conjunction and disjunction (i.e. 'and' and 'or') respectively. But they cannot be finitely defined in terms of conjunction and disjunction, because they make sense even if there are infinitely many objects. In that case, 'all' and 'some' behave like infinitely long conjunction and disjunction.

This analogy might make the quantifiers seem a small addition to the Boolean operations. But they are significant in two ways. First, there are striking examples of a subject-matter being rigourously shown to be part of another, in the above sense—once we allow quantifiers; with the attendant deduction of its doctrines from those of the other. The outstanding example is the demonstration—usually put in terms of deducing doctrines, rather than compounding properties—that all of classical pure mathematics is part of the theory of sets. This was a remarkable achievement, taking a half-century of effort (from about 1860 to 1910) by many mathematicians.

Second, infinitely long conjunction and disjunction suggest the more general idea of iterating any operation, infinitely rather than finitely. And this is significant for us. For take any well-defined collection of operations for compounding properties. (There are indeed others, not definable just in terms of the Boolean operations and quantifiers; e.g. interpolation in a spectrum of properties, or extension of such a spectrum.) Then there are clearly two main ways to interpret the above definition of one subject-matter being part of another; as follows.

Either one restricts oneself to finite iterations. This I call 'reduction'; though as noted above, this term has many other (and vaguer!) uses. This is of course the way you read the above definition: and it applies to the two examples above, squares and rectangles on one hand, and pure mathematics and set theory on the other.

Or one allows infinite, as well as finite, iterations. That is mind-stretching and requires care, if paradoxes are to be avoided: but it turns out to be tractable. We say 'infinitary' for 'infinite or finite'; so that this interpretation of the definition is sometimes called 'infinitary reduction' (just as the corresponding branch of logic is called ‘infinitary logic’). It is important because it turns out to be equivalent to a notion which is much discussed—so called 'supervenience'.

The exact definitions of 'supervenience' vary from one author to another. But the common, main idea is that one family of properties supervenes on another, with respect to a given set of objects (on which both families are defined), iff:

> any two objects that match for all properties in the second family (i.e. both having or both lacking each such property), also match for all properties in the first family.

Or equivalently, the contrapositive formulation: any two objects that differ in a property in the first family (one object having the property, the other lacking it) must also differ in some property or other in the second family. It is straightforward to show that these definitions are equivalent to infinitary reduction, as defined above.

A standard, largely uncontroversial example of supervenience is given by paintings. Most agree that the family of aesthetic properties of paintings (properties such as 'is well-composed' and 'has the colouring of a Matisse', or even the hack example, 'is beautiful') supervenes on their pictorial properties; where by 'pictorial properties' I mean some set of non-evaluative properties of very small regions like ‘is magenta for the one square milliimetre in top left corner’. Of course, if supervenience is to hold good, the family of pictorial properties must be suitably rich. In particular, they cannot just describe colour, even in a technical vocabulary, millimeter by millimetre; they must also describe details of the medium, e.g. oil or watercolour. But most would say that a family of non-evaluative properties can be picked out such that if two paintings matched utterly in respect of these properties, then (no matter how else they differ—maybe one is the original and the other is a fake!), they match in their aesthetic properties: if one is well-composed, so is the other, and so on.

This example also illustrates the contrast with (finite) reduction. Is a property such as ‘is well-composed’ a finitely long compound, built out of the pictorial properties? Many who are happy to accept supervenience have thought not: they point out that we hardly know how to begin writing such a finite definition, let alone how to improve it and perfect it. But here a warning is in order. Since a finite definition of ‘is well-composed’ could be so long as to be incomprehensible by human minds, even working collaboratively (e.g. a million million pages), our being unable to begin writing such a definition is very weak evidence for its non-existence. So reduction might yet hold.

3.C: Formulating Physicalism

We can now move towards formulating physicalism. For clarity, I shall first state the main idea, and then make it precise. The main idea will of course be that the mental reduces to, or at least supervenes on, the physical, in the above sense. That is, for reduction: each mental property is coextensive with some physical property, one of the simple given ones or a finite compound. For example, imagine that there is a compound (finite but perhaps very complex) physical property X such that: any person or animal is an instance of 'sees yellow' if and only if they are an instance of X. Then 'sees yellow' is reduced.

And the main idea of supervenience is: for any two objects, if they match on all their physical properties (i.e. for any physical property, they both have the property or they both lack it), then they match on all their mental properties. For example, consider a person during two seconds, who sees yellow. Imagine a physical duplicate (replica) of that person: this means that corresponding physical parts of the person and of the duplicate, right down to the smallest parts (the level of atoms or electrons or smaller still), are to be in identical physical states. And imagine that we control the environment of the person and of the duplicate, so that the duplicate remains a duplicate, moment by moment, for two seconds. (No doubt, to do this we will have to make the duplicate's immediate environment be itself a duplicate of the original person's immediate environment—a hard assignment!) Then supervenience implies: the duplicate also sees yellow.

So far as I can tell, most people, when asked whether such a duplicate would see yellow, confidently answer Yes. I confess: being a philosopher rather than an empirical sociologist, my evidence is anecdotal—when I say 'most people', I mean, strictly speaking, 'most philosophy undergraduates whom I have asked in a classroom survey'! But I ask them this question very early in the course, before I make any attempt to persuade them that physicalism is true. So though anecdotal, this evidence supports the point at the beginning of this Section: that physicalism is widespread in our intellectual culture.

On the other hand, these students' opinions are more divided, and of course more tentative, about whether reduction holds, i.e. whether there is a finite definition of ‘sees yellow’ in terms of physical properties. But the above warning is again in order. Since a finite definition of ‘sees yellow’ could be so long as to be incomprehensible, our inability to write such a definition is very weak evidence for its non-existence: reduction might yet hold.

So much for the main idea. To formulate physicalism precisely, we need to decide on two points: exactly what is the set of physical objects, of which the sets for other subject-matters will be a subset (cf. clause (i) of the definition above); and exactly what compounding operations among properties to allow (clause (ii)). Almost all of the literature focusses on the first point, though the second is obviously just as important; but I shall follow the literature, since I have nothing useful to say about the second point.

At first sight, it seems that clause (i) will only require that all the objects on which mental properties are defined are objects on which physical properties are also defined. That is: the mental subject-matter's set of objects is to be a subset of the physical subject-matter's set. And this seems uncontroversial. For we normally think of mental properties as defined on people and animals; for example, we say 'Fido sees yellow'; and people and animals also have physical properties!

But this clause is not so straightforward, for three reasons. The first two are closely related to the discussion in Section 2.B; the third is unrelated to previous discussion, and will lead us to the next Subsection.

(1) First: two subject-matters, say the physical and the mental, are typically presented to us with properties that do not have common instances. For example, '... sees yellow' has Fido as an

instance, even when we interpret '...sees yellow' as expressing a narrow mental state (cf. Section 2.B). But the physical properties that a physicalist claims to 'underly' this narrow mental state are typically presented as properties, not of dogs, but of brains or parts of brains: the standard example, in the philosophers' pretend-technical jargon, is '...is a firing of C-fibres'.

So to satisfy clause (i), we must expect to have to alter somewhat the properties (and so also the predicates expressing them) that are presented to us. In the example, either: we must introduce a mental property, corresponding to '...sees yellow', defined on brains or their parts, as it might be '...is a part of a brain in the narrow state of seeing yellow'; or we must introduce a physical property, corresponding to '...is a firing of C-fibres', defined on organisms, as it might be '...has a part of its brain that is firing its C-fibres'; or we can of course do both. To sum up: to satisfy clause (i), we must expect to do two things: make our set of objects contain parts of its own members, and/or wholes composed of its members; and 'massage' the properties accordingly.

(2) Second: we said at the start of this Section that physicalism claims *all* subject-matters, not just the mental, to be part of the physical. So clause (i) will require that all the objects on which *any* properties (e.g. biological or social properties) are defined are objects on which physical properties are also defined, i.e. are members of the physical subject-matter's set of objects. Indeed, even if one took physicalism to claim only that the mental is part of the physical, the social nature of mind, emphasised in Section 2, would prompt physicalism to put social objects, such as linguistic communities and nations, into the physical subject-matter's set.

This of course leads us back to the first reason, just above. The physicalist must somehow conceive objects like linguistic communities as physical (quite apart from issues about their properties). The obvious strategy for the physicalist is as just above. To spell it out: first, assume set-theory (or a similar device such as mereology, the theory of parts and wholes); second, conceive these objects as sets (or mereological wholes) with smaller, more obviously physical, objects—such as organisms or limbs or organs or cells or molecules—as members (or parts); third, take the set of physical objects as containing all sets (or wholes) composed out of objects that are given as being physical.

(3) The third reason for care about clause (i) leads us to the topic of modality, i.e. the notions of possibility and necessity. Here I shall show why we have to deal with these notions. In the next subsection, I will adopt a widespread, though admittedly controversial, way of doing so.

Clause (i) says 'a set of objects'. One naturally reads that as a set of actually existing objects, with their actual properties. But if we only consider the actual world (i.e. the actual total history of the universe, throughout all time and space), there is a grave risk that reduction—and even more risk that supervenience—will be trivially true: and true for a reason that has nothing to do with the intuitive idea of physicalism. So there is a risk that our formulations will not capture this intuitive idea.

The obstacle is very simple. Here is an example for reduction of a mental property. Suppose that as it happens, the set of actual instances of 'sees yellow' is finite. That is: throughout the actual world (in the above sense), only finitely many people and animals see yellow. Then there surely is a compound physical property with exactly that set as its instances. To find such a property, I could find for each instance of 'sees yellow', a physical property possessed only by that instance; and then form the disjunction of those properties. This finite disjunction will have exactly the yellow-see-ers as instances. Then 'sees yellow' is reduced. (For simplicity, I have stated the problem in terms of whole organisms as instances, and regardless of the issue of change, i.e. seeing yellow at one time but not another. The problem is unaffected by these points.)

Here is an example for supervenience. Suppose that no two actual people or animals exactly match in their physical properties. (That seems very likely!) Then supervenience is trivially true, on the usual understanding of 'all' and 'if, then'; as logicians put it, it is 'vacuously true'. (Again, the problem is unaffected by my choosing organisms as instances, and by my ignoring change.)

But the truth of physicalism—whether it be reduction, or supervenience—must not be established by the mere accident of the reduced property having finitely many instances, or of no two objects perfectly matching in the 'subvening' properties. So something has gone wrong with our formulation. (Being true on account of such a mere accident is of course sufficient for being contingent, which Section 3.A urged *was* a desideratum in formulating physicalism. But this desideratum should not be earned so cheaply: the accident of finitude or lack of match is clearly irrelevant to the intuitive idea of physicalism.)

3.D: The Range of Supervenience: Going Beyond the Actual World

We need a formulation of physicalism that somehow circumvents the mere accidents seen at the end of Section 3.C. That must mean: a formulation that is actually true, according to the physicalist, on account of how mental etc. properties would be reduced, even if there were infinitely many instances of them (or would supervene, even if there were perfect matches in physical properties). Such talk of how things would be, in contrary-to-fact circumstances, is often treated in terms of possible worlds. I will follow this tradition.

We can introduce possible worlds as follows. Many true propositions are contingently, not necessarily, true. The actual world makes the proposition true; but as we say, 'it did not have to be true'. Following Leibniz, and modern modal logic, we say: there is a possible world (different from the actual world) at which the proposition is false. So we imagine the set of all logically possible worlds. The actual world is one of them. They all make true all necessarily true propositions: the propositions of logic, maths, and analytic propositions such as 'All batchelors are unmarried'. And so also, in general: for any proposition there is a set of possible worlds at which it is true. (Any contradiction, or more generally impossible proposition, is associated with the empty set of worlds.) And any two logically equivalent propositions are associated with the same set of worlds. (For details of possible worlds' uses in philosophy, and debate about exactly what they are, see Kripke 1980, Lewis 1986a.)

Given this framework of possible worlds, reduction or supervenience will take as the set of objects a set of actual and possible objects. The absorption of other subject-matters by the physical, that reduction and supervenience claim, will be logically stronger: because holding on a larger set. Example: even if no two actual people or animals exactly match in their physical properties, surely there could be a person who exactly matches my present physical properties. So there is a possible world containing such a person. So if supervenience takes such a person in such a world, together with me, as members of its set of objects, then: supervenience requires that I and that person match in our mental etc. properties.

Here I have again taken an example with organisms as the objects (instances of the property) in question. I will continue to concentrate on this case. But as above, my discussion will be unaffected by this choice. We could accommodate the 'width' of the mental and social by having large regions of spacetime—or even entire possible worlds—as the objects. With these alternatives, authors often speak of 'regional', and 'global', supervenience, respectively; cf. Kim (1984, p. 168), Horgan (1982, pp. 36-37).

So the question is: exactly which set should be taken as the set of objects? Presumably, both reduction and supervenience should take, for any given possible world, either all the people and animals in that world, or none of them. (And similarly for organisms' parts, and wholes composed of many organisms; cf. Section 3.C.) So: which worlds should be considered?

This is a hard question, whose resolution depends on deep and controversial issues unrelated to mind: for example, issues about laws of nature, and the identity-conditions for properties. I can do no more than broach the issues. I shall do this by discussing two simple choices of worlds: choices that relate to the discussion above, and to the quantum theory below.

(a) Choose the set of all logically possible worlds. With this choice, reduction would be a matter of necessary coextension of properties. Supervenience would be a matter of: any two logically possible items that physically match also match mentally and in all other respects.

I reject choice (a), since it makes physicalism either logically necessary or logically impossible. (For it makes physicalism a matter of a pattern across all the worlds, and thereby true at all worlds or at none.) And I said in Section 3.A that I believed physicalism should be formualted so as to be contingent. But I should remark that some authors are content with choice (a), and so endorse physicalism as necessary.[5]

There is another general advantage in rejecting choice (a). It concerns the identity of properties. Namely: rejecting (a) gives a physicalist, even a believer in reduction, ample scope to agree with non-physicalists that mental, social etc properties are *not* identical with physical ones. The reason is that most philosophers agree that for two properties to be identical, they must be necessarily co-extensive. At least: almost all philosophers who are willing to talk at all about identity-conditions for properties agree on this; (they then dispute whether necessary co-extension is also a sufficient condition for identity of properties). If we agree with this, then a physicalist who rejects (a) is accepting that mental etc. properties are not identical with physical ones; for a necessary condition of such identity is violated.

This advantage of rejecting (a) is uncontentious. But it is worth emphasising, for two reasons. First, it may sugar the pill of physicalism for some who find it incredible that a mental property could be a physical property. (Lipton (this volume, Section 2) is such a person.) The sugar is that they do not have to believe this: physicalism only requires a contingent coextension—albeit over some decently wide class of worlds (yet to be specified!), so as to avoid the risk of merely accidental truth.[6]

Second, this advantage of rejecting (a) may seem to conflict with my own talk about properties being identical when they are co-extensive. But recall my remarks at the start of Section 3.B: I warned that my usage was convenient but unusual—and was not intended to deny the standard view that contingently co-extensive properties are not identical.

(b) The second choice for the set of worlds invokes the idea of a law of nature. Many philosophers agree that the actual world has laws of nature: and that these are, or at least correspond to, contingently true universal generalizations. For us, nothing turns on the difference between ‘are’ and ‘correspond to’. The main point is that laws differ from the merely accidentally true universal generalizations, e.g. 'the coins in my pocket are silver', by being in some way very informative about the world; though the exact nature of this informativeness is hard to analyse, and controversial.

Given this notion of law (however ‘informative’ is analysed), there is a well-defined set of all worlds that make true all the actual world's laws of nature. It is a subset of the logically possible worlds, since laws are contingent; a subset which includes the actual world—since the laws are actually true! This set is often called the (set of) nomically possible worlds (again, after the Greek word, νομοσ, for ‘law’).

I shall take this set as the second choice that physicalism might make. So reduction would then be a matter of nomic coextension of properties. Supervenience would then be a matter of: any two nomically possible objects that physically match also match mentally (and in all other respects).

As is obvious from the discussion of (a) above, this choice has the merit of having physicalism avoid claims of property-identity. It also makes physicalism contingent, whether as reduction or supervenience. This arises from each law being a contingent proposition (and indeed on some theories, a law might be a true proposition at a given world, and yet not be a law there). So different worlds can have different collections of laws; so there may well be logically possible worlds where the laws allow physical duplicates to differ mentally—say by having mental laws that are unrelated to the physical laws.

I believe that physicalism, once formulated as supervenience with this choice of worlds, is plausible. In particular, I believe it can accommodate the peculiar features of consciousness; pre-eminently, conscious states' intentionality (their representing objects and/or states of affairs) and their subjectivity (their involving 'phenomenal qualities', or 'qualia').[7]

But I cannot enter details. I have no space; and by and large, I have nothing to add to excellent discussions elsewhere. But let me single out Dennett's (1991, Chap. 12) and Lewis' (1990) arguments that physicalism can accommodate 'qualia'. Although Dennett does not defend a precise formulation of physicalism, let alone one involving possible worlds, Lewis of course does. He favours a definition of physicalism close to choice (b) above. The principal difference is that he uses his theory of natural properties; he also copes with the 'width' of the mental and social by taking entire worlds as the objects. (For both differences, cf. (M5) on p. 364 of his (1983), while choice (b) corresponds to his (M4)). Despite this difference, Lewis' argument about qualia can be easily adapted to choice (b): his central idea— that knowing what an experience is like is a kind of knowing how, not knowing that—is unaffected by the difference.

However, there is a problem with choice (b). Although I have no space to discuss it in detail, I should describe it. For in Section 5.B, I shall show how it is illustrated by an interpretation of quantum theory—giving my second conclusion, as announced in Section 1.

The problem is that a world where intuitively physicalism is false, because there are mental properties that are intuitively not supervenient nor reducible, may yet be a world where my formulation, with choice (b), is true—because the mental properties are correlated with physical properties according to strict (i.e. exceptionless) laws of nature. That is: choice (b) seems too weak; physicalism so formulated is too easily made true.

This problem is well-known in philosophy. (It is often noted in discussions of specific physicalist theories such as mind-brain identity theory and functionalism). There are two broad strategies for solving it. The first is conservative: keep the idea of formulating physicalism as a supervenience thesis, but make a different choice of worlds, for example in the way proposed by Lewis (ibid.). The second is radical: give up supervenience and try to make sense, in some quite different way, of the idea that mental properties are 'constituted' by physical properties. In the literature, this second strategy is gaining ground. But there is much disagreement—and I think, vagueness— about this new sense of 'constitution'. (Crane (1995, p. 212-213) gives references for this strategy; he also argues that this strategy cannot accommodate mental causation.) Hence my preference for the first strategy. But I cannot justify this here: it must suffice to notice the problem, and the two strategies.

So much by way of formulating physicalism. Before turning to more details about physics, let me summarize the story so far by emphasising the four main reasons why physicalism, as I have formulated it, is not 'reductionist' or 'eliminativist' about the mind (or indeed, any other subject-matter!).

(1) Reduction and supervenience, in my senses, clearly legitimate, rather than eliminate, the reduced or supervening subject-matter—squares, and talk of them, are legitimated, not eliminated, by being shown to be part of the subject-matter, rectangles.

(2) For a reduction in my sense to be useful for science, it must be manageably short. But this discussion, tilted as it is to metaphysics, has allowed a reduction to be arbitrarily long.

(3) Many believe that for reductions (in any sense) to be eliminative and/or useful, the reducing theory (or subject-matter or what-not) must provide explanations of the phenomena that are treated by the reduced theory. Maybe so: but my physicalism is not committed to such physical explanations of mental phenomena. For it is clearly compatible with pluralism about explanation (*a la* Lipton, this volume Section 3)—a pluralism which I in fact endorse.

(4) Physicalism does not require the identity of mental with physical properties, at least in the usual sense that involves necessary coextension. It requires only coextension across a suitably large class of possible worlds, including the actual world.

4. Quantum Theory and its Measurement Problem

But physicalism is only as precise as the word 'physical'! At the end of Section 3.A, I discussed how to make this word, and especially 'physical property', precise in terms of similarity to the known physical properties. But however successful that strategy might be for formulating physicalism, it leaves untouched two dangers, which I will treat in two subsections. The first, lesser danger concerns why people believe physicalism is true. The second danger concerns its being true, and is our main business. It arises from quantun theory's measurement problem.

4.A: Naivete about Physics

The first danger arises from a cluster of views, widespread in our intellectual culture, about physics as a science: that it has as its subject-matter, matter in motion, which it describes with precise mathematics; that what it tells us about this subject-matter is cumulative (i.e. it never gives up previously established claims); and even that at any given time, discussion among practitioners is uncontroversial. In short, the picture is of physics as a concrete floor of established, precise facts about simple concepts of matter and motion: a floor so firm (albeit perhaps dull!) that other sciences can build upon it. Needless to say, this picture is false. Physics is much more interesting than this picture suggests! It has a much more varied, and strange, subject-matter than the matter in motion of classical mechanics; and it is steeped in controversy. (For an antidote to this false picture, cf. Leggett 1987, esp. Chap.s 5 and 6.)

But though false, this picture may well influence people (or at least my philosophy undergraduates!) to think that physicalism is true, at least in the sense of supervenience. Here I have in mind a general, and a specific, point. The general point is that the usual verdict, in the thought-experiment about the physical duplicate of the person seeing yellow—that the duplicate also sees yellow—may well be influenced by the picture of people (and other organisms), or more specifically their brains, as composed of countless little particles.

The specific point is that a traditional argument against interactionism is flawed, because of this false picture of physics. Interactionism is the view, mentioned in Section 2.A, that mind and matter are connected by causation and/or law, but neither is reduced to the other. The traditional argument against it has various forms; but it is often presented in terms of energy. The idea is that any causal interaction between mind and matter would violate the principle of the conservation of energy. Thus, if irreducible mental properties (or states, events or what-not) caused physical ones—for example in wilfully raising my arm—that would surely mean that energy would flow into the realm of the physical. Similarly for physical properties or states causing irreducible mental ones: surely energy would flow out of the physical realm. But, says the argument, physics tells us that energy is conserved in the sense that the energy of an isolated system is constant, neither increasing nor decreasing. When it seems to change, the system is in fact not isolated, but rather gains energy from its environment, or loses energy to it. And there is no evidence of such energy gains or losses in brains. So much the worse, it seems, for interactionism. (Though traditional, the argument is still current; for example, Dennett endorses it (1991, pp. 34-35).)

This argument is flawed, for two reasons. The first reason is obvious: who knows how small, or in some other way hard to measure, these energy gains or losses in brains might be? Agreed, this reason is weak: clearly, the onus is on the interactionist to argue that they could be small, and indeed are likely to be small. But the second reason is more interesting, and returns us to the danger of assuming that physics is cumulative. Namely: the principle of the conservation of energy is not sacrosanct. The principle was only formulated in the mid-nineteenth century; and although no violations have been established hitherto, it has been seriously questioned on several occasions. It was questioned twice at the inception of quantum theory (viz. the Bohr-Kramers-Slater theory, and the discovery of the neutrino). And furthermore, it is not obeyed by a current proposal relevant to us (which we will discuss in Section 5.A): a proposal for solving quantum theory's measurement problem.

In short: physicalists need to be wary of bad reasons to think physicalism is true, arising from naivety about physics.

4.B: Avoiding an Indefinite Macrorealm

I turn at last to quantum theory (from here on, QT). As I said in Section 1, QT has an interpretative problem, called the measurement problem. There are many different strategies for solving this problem, but several of them are relevant to the mind-matter relation—more specifically, to physicalism. The rest of this Section will sketch the measurement problem, and how it bears on the mind-matter relation. Section 5 will take up some strategies for solving it.

Roughly speaking, the measurement problem is: QT's laws about how the states of objects change over time seem committed to the prediction that macroscopic objects often have no definite positions—nor definite values for other familiar physical quantities like momentum or energy. ('Quantity' is jargon for 'numerically measurable property'; 'magnitude', 'variable' and 'coordinate' are also used.) But this seems manifestly false: tables and chairs surely have definite positions etc. As it is sometimes put: the macrorealm is definite. (Or at least, we experience the macrorealm as being definite. So, if QT is to account for our experience, it must either secure such definiteness, or at least explain the appearance of it. But I postpone till Section 4.C this second strategy, i.e. the idea of allowing an indefinite macrorealm and securing only definite appearances.)

This problem is called the ‘measurement problem’, mainly because the argument that QT implies an indefinite macrorealm is clearest for a measurement situation. For QT says that microsystems, like electrons and atoms, in general do not have definite values for quantities. And if you use QT to analyse a measurement of, say, the momentum of an electron, which QT says has no definite momentum, you find that according to QT, the indefiniteness of the electron’s momentum is transmitted to the apparatus’ pointer—so that it has no definite position. I turn to spelling this out.

Like any physical theory, QT assigns states to systems: the state fully specifies the properties of the system. ('System' is just jargon for 'object'.) But the orthodox interpretation of a quantum state is as a catalogue of probabilistic dispositions. That is: for each quantity (position, energy, momentum etc.), the state defines a probability distribution on all possible values of the quantity. These states are represented by vectors: they are often written (in Dirac’s notation) with angle-brackets, e.g. | >. For each state, there are some physical quantities and some value of each such quantity, such that: the state ascribes probability 1 to that value for that quantity. The state is an ‘eigenstate’ of the quantity, the value an ‘eigenvalue’. But for each state, the great majority of quantities are ascribed a non-trivial probability distribution. This distribution is coded in the geometry of the vector space: the state is a vector sum of the quantity’s eigenstates. And it is called a ‘superposition’ and is written with a '+'.

So far, so good. But the orthodox interpretation of QT adds that the system has a value for a given quantity *only when* its state ascribes probability 1 to that value. This is called the ‘eigenvalue-eigenstate link’.[8] It is this scarcity of values that leads to the measurement problem. For the interaction of, say, an electron that is in a superposition (not an eigenstate) for momentum, with an apparatus for measuring momentum, leads to the electron’s indefiniteness being transmitted to the apparatus—so that its pointer is in no definite position! This suggestion has been proven for a wide range of exact quantum theoretic models of measurement. But we can confine ourselves to a very simple model.

As an example, I will take a momentum measurement on an electron in a superposition of two momentum eigenstates: one for 1 unit of momentum, and the other for 2 units of momentum. Suppose we have a measurement apparatus or pointer, with ‘ready state’ |r >, which reliably reads these eigenstates, in the sense that the composite system behaves as follows:

| 1>|r > —> |1 >|reads ‘1’ > and | 2>|r > —> |2 >|reads ‘2’ >.

Here, the juxtaposition of two kets represents a state of a composite system, in our case the electron+pointer: namely the 'conjunction' of the two juxtaposed states. And the arrow represents the evolution of the state in time, as prescribed by QT's famous Schrödinger equation. So each of these displayed formulas means: if the composite electron+pointer is begun in the state on the left, then it evolves by the Schrödinger equation in some fixed finite time to the state on the right.

Then the Schrödinger equation (which is the principal law of QT) implies that measuring an electron initially prepared in a superposition yields:

$$\{| 1> + |2 >\}|r > \longrightarrow |1 >|\text{reads ‘1’} > + |2 >|\text{reads ‘2’} >.$$

But the final state on the right is not an eigenstate of position for the pointer. So the orthodox interpretation of QT, more precisely the eigenvalue-eigenstate link, is committed to the pointer having no definite position!

There are clearly two main approaches for solving this problem. Either we somehow change the Schrödinger equation, so as to replace the above final state by an eigenstate of pointer-position. This approach is called 'collapsing the wave-packet'. I discuss it in Section 5. Or we somehow supplement the eigenvalue-eigenstate link’s meagre ascription of values: we postulate extra values. But in this paper, I have no space to discuss this second approach. Suffice it to say that like Section 5's approach, it has versions that uphold physicalism, and versions that violate it; (cf. my 1995, p. 145f; 1996).

4.C: Avoiding Indefinite Appearances

For our topic of QT and mind, it is also important to emphasise another contrast (briefly mentioned at the start of Section 4.B). Namely, between:

(DefMac): those strategies for solving the measurement problem that aim to secure a definite macrorealm, and so to explain why the macrorealm is as it appears to be; and

(DefApp): those, perhaps more radical, strategies that allow an indefinite macrorealm and aim only to secure definite appearances (thus denying that it is as it appears to be).

We shall see in Section 5 that this contrast cuts across the one at the end of Section 4.B. That is: some versions of the 'collapsing the wave packet' approach aim to secure a definite macrorealm, while others aim only to secure definite appearances. (And similarly some versions of the 'extra values' approach aim for a definite macrorealm, e.g. the Bohm interpretation; while others aim only for definite appearances, e.g. the 'many minds' interpretation.)

To clarify the contrast between (DefMac) and (DefApp), it will be helpful to 'drive the measurement problem into the brain'. That is: it will be helpful to show how, according to QT's orthodox interpretation, a quantum-theoretic model of perception of the pointer will lead to an indefinite perception of pointer-position, in the case where the electron is initially prepared in a superposition. Since we never have such indefinite perceptions, orthodoxy will face a problem of 'indefinite appearances', just as much as one of 'indefinite macrorealm'. Showing this will occupy the rest of this Section.

We can again consider a very simple model of perception, based on our toy-model of measuring an electron's momentum. Recall that we had:

$$\{| 1> + |2 >\}|r > \longrightarrow |1 >|\text{reads ‘1’} > + |2 >|\text{reads ‘2’} >.$$

Now let us assume that the brain-state corresponding to a person, Anna, observing the pointer ‘s position, and believing it to be at ‘1 unit’, is a quantum-state, call it

| believes ‘1’>. And similarly for her brain-state corresponding to observing and believing it to be ‘2 units’: it is a quantum state, say | believes ‘2’>. In other words, let us assume a physical quantity for Anna’s brain, with eigenvalues 1, 2 etc., whose eigenstates correspond to her beliefs in such a way that we can mnemonically call it 'belief-in-position'. (Despite the mnemonic, it is of course a normal physical quantity, perhaps some function of the positions, momenta, energies etc. of her brain's constituent particles). And similarly for her brain-state corresponding to her ‘ready state’, | alert> say.

These assumptions countenance using QT to describe the brain. That is a substantial contention, based of course on taking QT to be the fundamental theory of all (material, spatiotemporal) objects. (For further defence of these assumptions, see e.g. my (1995), pp. 146-147.) But given these assumptions, we can 'drive the measurement problem into the brain'. In terms of our toy-model: assuming that Anna is reliable on the eigenstates, in the sense:

| 1>|r >| alert> —> |1 >|reads ‘1’ >| believes ‘1’> and:
| 2>|r >| alert> —> |2 >|reads ‘2’ >| believes ‘2’>,

the Schrödinger equation implies that measuring a superposition gives:

{| 1> + |2 >}|r >| alert> —>
{|1 >|reads ‘1’ >| believes ‘1’> + |2 >|reads ‘2’ >| believes ‘2’>}

which is not an eigenstate of pointer-position, nor of the quantity we called 'belief-in-position'! But Anna—we!—always have definite beliefs about pointer-positions. So (given the orthodox eigenvalue-eigenstate link), the fact that the final state is not an eigenstate of belief-in-position seems manifestly wrong.

To sum up:— We have seen that QT faces a problem of 'indefinite appearances', just as much as one of 'indefinite macrorealm'. And there are two broad strategies for responding to these problems, viz. (DefMac) and (DefApp) above. That is, we can either:

(DefMac): somehow secure a definite macrorealm; and more specifically secure the successful predictions of classical physics, so that we can rely on a classical psychophysics for understanding the definiteness of appearances;

or we can:

(DefApp): allow an indefinite macrorealm, and somehow secure only that appearances are definite.

In view of the discussion above, we expect that (DefApp) will involve some 'funny business' at the interface of brain and mind. And indeed, it is exactly here, under strategy (DefApp), that some proposals violate physicalism.

5. Collapsing the Wave Packet

This Section discusses one approach to the measurement problem: the approach that postulates new physical laws to replace the Schrödinger equation. Such a change of the state is called 'the collapse of the wave-packet' (or 'state reduction'). More specifically, I take three versions of this approach. In Section 5.A, I take that of Ghirardi, Rimini, Weber, Pearle et al.; and that of Penrose. Section 5.B discusses that of Wigner and Stapp. For all versions, I emphasise their consequences for brain or mind, and so for physicalism; (I take them in increasing order of radicalism).

The first two versions (Section 5.A) take the collapse of the wave packet to be an purely physical process; and so aim to secure a definite macrorealm (i.e. strategy (DefMac) of Section 4.B). Perhaps unsurprisingly, the new laws describing the collapse of the wave packet seem entirely compatible with physicalism as defined in Section 3. But these versions substantiate my first conclusion, announced in Section 1: on these proposals, the *way* in which the mental supervenes on the physical would be very different from what most people, unversed in the controversies about QT, would expect. (The difference is more radical, for Penrose's version.)

On the other hand, Wigner and Stapp's version invokes mind to trigger the collapse of the wave packet. This will yield my second conclusion: namely, this version gives a real-life example of the abstract problem confronting the definition of physicalism at the end of Section 3. So here is an interpretation of QT that violates the spirit of physicalism, if not its letter.

5.A: Physical Collapse

I will first discuss the proposals of such authors as Ghirardi, Rimini, Weber, Pearle and Percival (giving me the acronym GRWP); who propose precise equations for the collapse of the wave packet (to one out of various alternative eigenstates). They do not aim to discuss the brain. But it turns out, surprisingly, that according to their proposals, the collapse sometimes happens, not in the external world, but in the observer’s nerves (e.g. the retina). This is no threat to physicalism, but it *is* surprising.

I shall concentrate on a well-known precise proposal, called 'continuous spontaneous localization' ('CSL': Ghirardi et al. 1986, Pearle 1989). This postulates a jitter that continually gives ‘little hits’ to the quantum state, in addition to its usual evolution according to the Schrödinger equation. (These hits increase the system's energy: as mentioned at the end of Section 4.A.) The jitter is a stochastic process, which in any individual case has a realization, with a prescribed probability. (Analogy: The jitter is like the probability space for 10 tosses of a coin; the realization is a specific sequence of 10 results, heads or tails.) Which realization happens in an individual case determines what happens to the system—what the final eigenstate is.

Any such proposed dynamics must somehow make the collapse mechanism ineffective in the microrealm (so as to recover the empirical success of QT's Schrödinger equation), but effective in the macrorealm. To do this, CSL chooses a rate of hitting and a size of hits so that the effect is almost always utterly miniscule for a microsystem. But on the other hand, each individual component of a composite system is subject to these rare and weak hits; and when hit, it drags the other components with it. The result is that for a macro-system with maybe 10^{23} components, the collapse is very fast (10^{-9} seconds). To take the time-honoured example: Schrödinger's cat is only superposed between life and death for a split second! This is surely acceptable, even to an advocate of the strategy (DefMac): for though it only gets the macrorealm to be definite at almost all times, nobody can be so certain that it is definite *always*.

The further details of CSL need not concern us. The mere fact that it links collapse to having a large number of components is enough to yield interesting consequences for physicalism. The point is best made by presenting an objection to CSL, due to Albert and Vaidman (cf. Albert 1992, pp. 100-111). The objection is wrong, but fruitful: for the reply brings surprises about the way in which the mental supervenes on the physical.

The idea of the objection is that some measurement results involve only a small number of microsystems. For example, a result may be registered by the arrival of a microsystem in one position rather than another on a fluorescent screen (like a TV screen). But its arrival only excites about ten atoms, which then de-excite each emitting a photon. Since retinal cells are so sensitive as to fire in response to a few photons, this is enough for a human to detect one result rather than another. So the objection is that since, at any stage in this process, only a few microsystems are involved, there will only very rarely be a GRWP collapse -- contradicting the fact that our perceptions of the spots on the screen are always definite.

GRWP reply by analysing the physical process of a nerve cell firing. They show that the transport of ions involved in the firing requires sufficiently many particles being sufficiently well separated in space for a collapse to occur with overwhelming probability in, say, a hundredth of a second. (Cf e.g.Ghirardi et al. 1995 Section 5.2.)

I should make two points about this reply. First, it involves no appeal to mind or consciousness: and so it poses no threat to physicalism. For it rests on the purely physical and contingent fact that in the example, the nervous system is the first place where enough particles are involved for the theory to predict collapse. If our retinas responded only to millions of photons, or if we perceived a result only by reading words or by hearing (which both involve displacements of millions of atoms), all collapse relevant to our perception of a definite result would indeed occur outside the head.

This leads in to the second point. Namely, that such examples of collapse occurring only in the head, so late in the causal chain of perception—although in no way threatening physicalism—are indeed surprising. For GRWP's proposals are examples of strategy (DefMac) of Section 4.C: i.e. of aiming to secure a definite macrorealm, which obeys classical physics (to a very good approximation). And on this strategy, one naturally expects psychophysics (which is to then secure the definiteness of appearances) not to involve any peculiarities of quantum theory. After all, consider the success of neuroscience in understanding perception, while using a wholly classical chemistry: i.e. a chemistry which models a molecule as like a group of billiard balls linked by rods, not subject to such quantum peculiarities as being in a superposition of two positions. (Nor a superposition of other quantities, including ones especially relevant to chemistry, such as handedness: classical chemistry does not countenance superpositions of left-handed and right-handed orientations of a molecule!) This example, and others like it, lead one to expect all such quantum peculiarities to 'die out' at the level of neurophysiology—and so also at the level of psychophysics. (For a general discussion, bearing on reduction and cumulativism in science, cf. Rohrlich & Hardin 1983.) So it is surprising that GRWP's proposals entail that psychophysics involves such peculiarities.

To sum up: GRWP's proposals are compatible with physicalism, but hold some surprises about exactly how the mental supervenes on the physical.

I turn now to Penrose. I will argue that the same overall conclusion applies to him; though in his case, the surprises for psychophysics are greater. Unlike GRWP, he is uncommitted about precise equations; but he thinks gravity is responsible for the collapse. He also deliberately aims to discuss the brain, indeed the mind. For he believes that although the collapse is a purely physical process, it involves non-computational physics; and that this physics will be relevant to brain action, because it will help explain connsciousness, which Penrose believes to be non-algorithmic (non-computable). He even has proposals about how this physics operates in the microstructure of cells (1989, especially pp. 367-371; 1994, especially Chapter 6, section 8 et seq., and Chapter 7).

Unlike GRWP, Penrose does not yet have detailed models of how gravity induces collapse. But his idea is that the gravitational self-energy of the difference between two mass distributions considered to be in quantum superposition determines a rate at which collapse takes place to one of the two distributions (1994, Sections 6.10-6.12, especially 6.12). This idea is akin, as Penrose of course acknowledges, to GRWP's proposals; and this idea, suitably developed, seems quite as likely to solve the measurement problem satisfactorily, as are the proposals of GRWP.

I turn to Penrose's claim that consciousness is non-algorithmic. This claim is very controversial: it is largely based on an analysis of Gödel’s monumental 1931 theorem about the incompleteness of arithmetic. As Penrose discusses (1994, Part I), it is uncontroversial that Gödel’s theorem establishes that human mathematicians are not using an algorithm that is both sound and knowably so. He goes on to argue that they are not using an ‘unconscious algorithm’, i.e. one that is sound but not knowably so. If that is right, then at least one aspect of human consciousness—viz. mathematical understanding—would be non-algorithmic. Penrose goes on to suggest that other aspects of consciousness (with some other organisms included) are also non-algorithmic.

Penrose connects the idea that consciousness is non-algorithmic with collapse, by arguing that both classical physics and orthodox QT are, in the relevant senses, algorithmic. So he maintains that although collapse is a physical process happening all the time, outside the body, due to gravity (thus securing a definite macrorealm), it has a scientifically important role inside the brain: to supply the non-algorithmic physics that underlies consciousness—or at least mathematical understanding.

Penrose also proposes a biological locus for this non-algorithmic physics. After first conceding that neurons are ‘too large’ and ‘classical’ to be involved in this new physics (1994, sections 7.1-7.2), he proposes the microtubules that occur in cells. These tiny filaments have some promising features (1994, sections 7.3-7.7). In particular, they might allow an internal quantum state: (i) to be coherent, relatively unperturbed by the environment; and (ii) to interact with a classical computation performed along the surface of the microtubule, by the varying configurations of the tubulin protein molecules making up that surface (rather in the manner of cellular automata).

Turning to assessment, I obviously cannot do justice to this slate of imaginative, and mutually-connected, proposals. It must suffice to make two comments. First: my own reactions, for what they are worth. I am happy to allow that gravity is crucial to collapse of the wave packet, if such there be; and also that consciousness, or at least mathematical understanding, is non-algorithmic. But I am not convinced that both classical physics and orthodox QT are, in the relevant senses, algorithmic. So collapse may not be the only mechanism that could provide the non-algorithmic basis of mathematical understanding. And so microtubules may not be the only place to look for such a basis.[9]

Second: the overall conclusion for us is as it was in the case of GRWP. If Penrose's proposals are true, then: physicalism is intact, but psychophysics is full of surprises! As Penrose says: ‘it is only the arrogance of our present age that leads so many to believe that we now know all the basic principles that can underlie all the subtleties of biological action’ (1994, p.373).

5.B: Mental Collapse

I turn to the proposal of Wigner and more recently Stapp, that mind (or consciousness) itself produces the collapse of the wave-packet; which yields my second conclusion, as announced in Section 1.

The idea is that the usual Schrödinger evolution holds throughout the physical realm, and is broken only at the interface of brain and mind. Once the mind sees one result (in our example, once Anna sees one pointer-position), the superposition is replaced by an eigenstate, namely the one corresponding to the result seen.

The first point to notice about this proposal is that it is a hybrid of the broad strategies (DefMac) and (DefApp) in Section 4.D. For once consciousness has 'done its stuff' and collapsed the wave packet, the macrorealm really is definite (in the quantities of which the collapsed state is an eigenstate). But then the usual Schrödinger evolution takes over again. So in general, there is no guarantee that the macrorealm stays definite (in those favoured quantities) as times goes on, in particular when consciousness stops looking—'when there's no one about in the Quad'.

Among the founding fathers of QT, Wigner (1962) expresses this proposal most clearly; (although he later changed his mind, and there are similar views in von Neumann and Heisenberg). Stapp (1993) revives the proposal; of course acknowledging these precursors. Unsurprisingly, the authors differ about what exactly it takes to reduce the state. Thus Heisenberg allows collapses to occur not only in humans, but also in cats; (and even in inanimate macroscopic apparatuses). Stapp (1993) joins him in this, but has now withdrawn this allowance, on grounds of parsimony (1995).

There are two obvious questions confronting this proposal, as so far stated. I need to pose them, but not to answer them. For posing them will be enough to yield my conclusion: that this proposal violates the spirit, if not the letter, of physicalism as we formulated it in Section 3.

The first question is: with which quantity's eigenstates does consciousness replace the initial superposition? No doubt, the basic idea of the answer must be 'the quantity that seems to have definite values'. But then the question is: what general laws, if any, constrain which quantity that is? If there are such laws, can they be expressed in wholly physical terms; or is there some irreducible invocation of mind? At first sight, it seems that (1) if the laws are physical, then physicalism might yet be upheld on the Wigner-Stapp proposal; while (2) if they must invoke mind, then it cannot be upheld.

The second claim (2) is straightforward: no doubt, the existence of irreducibly mental laws implies that physicalism is false. But about the first claim, (1), matters are not so clear-cut. For we are back at the problem that confronted Section 3.D's formulation of physicalism (i.e. using choice (b): supervenience across the set of nomically possible worlds). That problem was: if mental properties are correlated with physical properties according to strict laws, then this formulation can be true, even though, intuitively, physicalism is false. (In other words: choice (b) seems too weak, and the range of worlds across which physicalism claims supervenience needs adjusting.).

This problem applies here too. Even if the laws (governing which quantity's eigenstates are collapsed onto) are expressed in wholly physical terms, the very fact that only with consciousness is there any collapse means that, intuitively, physicalism is false. Similarly, if there are no laws at all about which eigenstates the collapse is onto: intuitively, physicalism is false.

A similar point applies about the second obvious question confronting this proposal. This question concerns the time of collapse. Thus, we can ask: what if anything constrains or determines when the collapse occurs? Suppose that laws do so, and they can be expressed in wholly physical terms. Then nevertheless, just as before, intuitively physicalism is false—because only with consciousness is there any collapse. Similarly, if there are no laws at all about when collapse occurs: intuitively, physicalism is false.

I cannot here pursue how one might answer these two questions, nor how to improve the formulations of physicalism. Suffice it to say, by way of summary: here is a real-life, albeit rather undeveloped, proposal, which violates the intuitive idea of physicalism—and provides an example of a problem which is discussed only in the abstract in philosophy. One could therefore try to use the proposal as a test-case for conjectured precise formulations of physicalism.

Acknowledgements: I thank: Peter Lipton for comments on a draft of this paper; participants at the conference for discussions; Chris Daly, Sue James and Tim Crane for advice on references; Bay Whitaker for making a transcript; John Cornwell, Rebecca Beasley and the Science and Human Dimension Project, for the invitation—and for patience; and David Lewis for the inspiration of his metaphysical system, in writing Sections 2 and 3.

FOOTNOTES

[1] In any epoch since the time of Galileo, there have of course been processes or phenomena which seemed not to admit mechanical explanations, and for which some scientists accordingly hypothesised a non-mechanical explanation. But the definitive demise of mechanism within physics came as a result of the success of clearly non-mechanical theories; namely, theories of electromagnetism, around 1900.

[2] Furthermore, in Section 3 we will see that my formulation of physicalism dovetails neatly with this strategy of 'looking widely'.

[3] Though these states are hard to describe, for the reasons given, they can to some extent be conveyed to others—witness the works of the impressionists, and writings about the stream of consciousness, by authors such as Proust, Joyce and Woolf.

[4] For the prevalence, and legitimacy, of philosophy adding detail and precision to claims that are widespread in the contemporary culture, see Craig (1987). He gives telling case-studies, drawn from throughout the last 400 years.

[5] Teller (1984) says so explicitly; others are less explicit, e.g. McGinn (1991). But McGinn has recently (1996) been explicit in his criticism of Chalmers' (1995) claim that 'zombies'—physical duplicates of sentient beings like you and me, but lacking sentience—are logically possible.

[6] But I agree with Lipton that Searle goes wrong. Searle's analogy for making such a coextension credible, namely the relation of solidity to lattice-vibrations (1992, pp. 112-126; this volume, replies to Theses 7, 8, [=pp. 14-17 of MS].) has no relevant difference from examples, such as heat and molecular motion, which Searle claims to be false analogies.

[7] Indeed, I believe that a logically stronger doctrine about the mind-matter relation—namely, in the jargon, analytical functionalism combined with contingent type-type mind-brain identity theory—is plausible; and can accommodate these features of consciousness.

[8] Agreed, to a philosopher of probability, this identification of a fact (having a value) with its holding with probability 1, will seem a howler. And as we shall see, it may well be wrong.

[9] For a recent exchange emphasising biological details, cf. Grush & Churchland (1995), Penrose & Hameroff (1995).

REFERENCES

Albert D., 1992. *Quantum Mechanics and Experience*. Cambridge MA: Harvard University Press.

Baker G. & Morris K., 1993. 'Descartes Unlocked', *British Journal for the History of Philosophy* **1**, pp. 5-27.

Boden M., (this volume). 'Consciousness and Human Identity: an Interdisciplinary Perspective'.

Burge T., 1979 'Individualism and the Mental', *Midwest Studies in Philosophy*, vol. IV, ed. P A French et al.. Minneapolis: University of Minnesota Press.

Butterfield J., 1995. 'Worlds, Minds and Quanta', *Aristotelian Society Supplementary*

Volume **69**, pp. 113-158.

Butterfield J., 1996. 'Whither the Minds?'. *British Journal for the Philosophy of Science* **47**, 200-221.

Chalmers D., 1995. *The Conscious Mind*, Oxford: University Press.

Craig E., 1987. *The Mind of God and the Works of Man*, Oxford: University Press.

Crane T. & Mellor D., 1990. 'There is no Question of Physicalism', *Mind* **99**, pp. 185-206; reprinted in Mellor's *Matters of Metaphysics* (1991), Cambridge: University Press.

Crane T., 1995. 'The Mental Causation Debate', *Aristotelian Society Supplementary Volume* **69**, p. 211-236.

Dennett D., 1991. *Consciousness Explained*, London: Penguin.

Frisby J., 1979. *Seeing*. Oxford: University Press.

Ghirardi G., Rimini A. & Weber T., 1986. 'Unified Dynamics for Microscopic and Macroscopic Systems', *Physical Review* D**34**, pp. 470-491.

Ghirardi G., Grassi R. & Benatti F., 1995. 'Describing the Macroscopic World: Closing the Circle in the Dynamical Reduction Program', *Foundations of Physics* **25**, p. 5-40.

Grush R & Churchland P., 1995. 'Gaps in Penrose's Toilings', *Journal of Consciousness Studies* **2**, pp. 10-29

Healey R., 1978. 'Physicalist Imperialism', *Proceedings of the Aristotelian Society* **74**, 191-211.

Horgan T., 1982. 'Supervenience and Microphysics', *Pacific Philosophical Quarterly* **63**, pp. 29-43.

Kim 1984 'Concepts of Supervenience', *Philosophy and Phenomenological Research*, **45**, pp. 153-176; reprinted in Kim's *Supervenience and Mind* (1993), Cambridge: University Press.

Kripke S., 1979. 'A Puzzle about Belief', in *Meaning and Use* ed. A Margalit, Dordrecht: Reidel.

Kripke S., 1980. *Naming and Necessity*, Oxford: Blackwell.

Leggett A., 1987. *The Problems of Physics*, Oxford: University Press.

Lewis D., 1983. 'New Work for a Theory of Universals', *Australasian Journal of Philosophy* **61**, 343-377.

Lewis D., 1986. *Philosophical Papers* volume II, Oxford: University Press.

Lewis D., 1986a. *On the Plurality of Worlds*, Oxford: Blackwell.

Lewis D., 1990. 'What Experience Teaches', in W. Lycan ed. *Mind and Cognition*, Oxford: Blackwell.

Lipton P., (this volume). 'Binding the Mind'.

Locke J., 1972. *Essay concerning Human Understanding*, London: Dent Everyman

McGinn C., 1991. *The Problem of Consciousness*, Oxford: Blackwell.

McGinn C., 1996. Review of D. Chalmers' *The Conscious Mind*, in *Times Higher*

Education Supplement, April 5 1996, pp. vii-ix.

Midgley M., (this volume). 'Putting Ourselves together Again'.

Monod J., 1972. *Chance & Necessity* London: Collins.

Oppenheim P. & Putnam H., 1958 'The Unity of Science as a Working Hypothesis' in H Feigl, M Scriven and G Maxwell ed.s *Minnesota Studies in Philosophy of Science vol. 2*, Minneapolis: University of Minnesota Press

Pearle P., 1989. Combining Stochastic Dynamical State-Vector Reduction with Spontaneous Localization, *Physical Review* A**39**, pp. 2277-2292.

Penrose R., 1989. *The Emperor's New Mind*, Oxford: University Press.

Penrose R., 1994. *Shadows of the Mind*, Oxford: University Press.

Penrose R. & Hameroff S., 1995: 'What 'Gaps'?', *Journal of Consciousness Studies* **2**, pp. 99-112

Putnam H., 1975. 'The Meaning of 'Meaning'', in his *Mind Language and Reality*, Cambridge: University Press.

Rohrlich F. & Hardin C., 1983. 'Established Theories' *Philosophy of Science* **50**, pp. 603-617

Rose S., (this volume). 'The Rise of Neurogenetic Determinism'.

Searle J., (this volume). 'How to Study Consciousness Scientifically'.

Searle J., 1992. *The Rediscovery of Mind* Cambridge MA: MIT Press.

Stapp H., 1993. *Mind, Matter and Quantum Mechanics*, New York; Springer-Verlag.

Stapp H., 1995. 'The Integration of Mind into Physics', in Greenberger D. and A Zeilinger (ed.s), *Fundamental Problems in Quantum Theory*, New York:Annals of New York Academy of Sciences.

Teller P., 1984. 'A Poor Man's Guide to Supervenience and Determination', *Southern Journal of Philosophy*, supplement to vol **22**, pp. 137-162

Wigner E., 1962. Remarks on the Mind-Body Problem, in Good I.J. (ed.), *The Scientist Speculates*, London: Heinemann.