HUW PRICE

THE PHILOSOPHY AND PHYSICS OF AFFECTING THE PAST*

Most people working on the philosophical side of quantum mechanics will have heard of the suggestion that the theory should be interpreted as showing that at the quantum level, physical events may 'affect the past' (i.e., have earlier effects). Yet this approach to the conceptual problems has received little serious attention, even in comparison to views which themselves involve very profound conceptual revision, for example, Everett's [1957] 'many worlds' account. Backward causation seems to be regarded as intrinsically absurd, or paradoxical, even by the by-now taxed standards of the discipline. This is somewhat surprising, given that in 'mainstream' metaphysics, by contrast, the notion of backward causation is rarely seen as *patently* absurd. Its possibility is discussed quite often, and at length. Of those who come down against it, few would think the conclusion obvious.

This suggests to me a regrettable lack of communication between mainstream metaphysicians and the philosophers and physicists of quantum mechanics. It is increasingly unlikely that the problems of quantum mechanics will not require some major revision in our ordinary ideas of the world. There seems little hope of a speedy decision as to where this revision should be (or indeed as to how this should be decided) without the best efforts of philosophers, as well as physicists. It is surely up to philosophers to ensure that physicists do not ignore major and promising avenues, in the mistaken belief that these are patent philosophical dead ends.

Against this background, I here argue a philosophical case for the possibility of backward influence, and try to relate the discussion to the conceptual problems of quantum mechanics. Specifically, I argue that a conceptual scheme under which we would claim the ability to bring about certain past events is not only internally coherent, but a possible result of a modification of existing conceptual schemes in the light of experience (and therefore a live option in physics). I identify a crucial assumption, at present taken for granted, which such a revision would require to be given up. The conceptual consequences of such a move

Synthese **61** (1984) 299–323. 0039–7857/84/0613–0299 $02.50

would certainly be profound. But I think they might properly seem the least of the available evils.

I begin with an analysis of the general form of a claim to bring something about. This draws heavily on that of Dummett [1964]. Dummett sets out to discover the difference between our views of the past and the future which ordinarily leads us to say that we can only affect the latter. Unable to find an argument to show that the difference he identifies is a feature of any coherent conceptual scheme, Dummett concludes that affecting the past is a conceptual possibility. I concur, though I argue that Dummett misrepresents the accepted difference between past and future (perhaps as a result of a peculiarity of his main example). This might have obscured the relevance of his discussion to possible physical cases of backward influence. I defend my version of the analysis against a recent argument from D. H. Mellor. Mellor claims that Dummett's considerations in fact show the impossibility of backwards causation. Here, my response to Mellor serves to highlight the choice of conceptual evils which would be involved in accepting backwards influence.

I then turn to quantum mechanics. I indicate how the admission of backward influence would explain the peculiar phenomena from which the conceptual problems arise. And I compare more popular interpretations with positions which, in the earlier discussion, will have emerged as rivals to any claim to bring about the past. As in the general case, my conclusion is that the backward influence approach may well prove the most benign of a difficult bunch. As such, it deserves more sympathetic attention than it has received.

1.

When am I justified in claiming that I can *bring about* an event of a certain kind? Two things seem crucial: the *power* to act in a certain way; and the *sufficiency* of an action of this kind for the occurrence of the type of event in question. Formally, a claim that in circumstances of type C, I am able to bring about an event of kind E by means of an action of kind A, resolves into two subclaims:

(1) In circumstances C, action A would be sufficient to ensure event E.

(2) In circumstances C, it is in my power to perform an action A or not, as I choose.

To ensure that the claim is non-trivial, we should impose these conditions on C: the circumstances should not themselves be sufficient for E; and I should not have reason to believe that such circumstances will never arise.

This analysis is neutral as to whether E occurs before or after A. Yet it seems that we never admit such claims in the former case. Why should this be so?

The usual argument that it *must* be so runs like this. Suppose a person (George, to name him) did claim to be able to bring about an event E by means of an action A at a specified later time. Because the time for action is later than the time of the claimed effect, we can in principle find out whether E has occurred before George is required to act. This enables us to ask George to perform A if and only if the event E has *not* occurred. We can offer George any inducement, so we can suppose he tries to do as we ask. If he succeeds we have either A and not E, or E and not A. The former outcome refutes his claim (1) ('sufficiency'). The latter conflicts with the requirement that the prevailing circumstances not themselves be sufficient for E. On the other hand, if George cannot do as we ask, this refutes (2) ('power'). So whatever the outcome, we will have shown that George is not justified in claiming to be able to bring about E by means of A.

I shall call this the *causal loop argument* (*CLA*, for short). CLA embodies what I think are the two main sources of the feeling, widespread among physicists, that the notion of affecting the past is intrinsically absurd. One is the threat of paradox, made explicit in a common objection to the possibility of time travel: if I could travel to the past, then I could meet and kill my young self (thus disposing of his would-be murderer). The other is the feeling that it would conflict with free will to admit that our future decisions may already have present and past effects. CLA combines these intuitions, arguing that if future actions lead to past effects then either it is possible to generate paradoxical physical situations, or we must admit that such actions are not freely chosen (but rather themselves the effects of the events they are supposed to bring about). We shall see that as CLA is circumvented, both these fears are laid to rest.

2.

Dummett notes that CLA depends on a principle something like this: it

is in principle possible to know of the occurrence of any past event, independently of knowledge of one's own future actions.[1] Unless George accepts this, he may deny that it is possible to find out whether an event E has occurred, before (as he claims) he must perform A, in order to ensure E. If he does deny this, he will simply dismiss our thought experiment as impossible to perform, even in principle.

There is an analogous proposition about future events: it is in principle possible to know of the occurrence of a future event, independently of knowledge of one's own future actions. Dummett says that it is the fact that we ordinarily reject this principle which enables us to take ourselves to affect the future. The perceived difference between the past and the future in this respect thus results from our different attitudes to this pair of principles. Moreover, Dummett thinks it is conceivable that a person could reject the past-directed principle (as we reject the future-directed one) and hence claim the ability to affect the past.

Let us be clear about the intended sense of the term 'possible' in these two principles. The past principle might seem to depend on the claim that every event *actually* leaves its traces on all future times; that the evidence is there in the world, for anyone clever enough to decipher it. This 'archivalist' view of the past is perhaps more attractive, on the face of it, than the analogous view of the future. But our ability to affect the future but not the past cannot depend on this difference. Archivalism has very little of the powerful intuitive appeal of the principle that we can't influence the past. We don't take our present inability to affect say Nixon's involvement in the Watergate break-in to rest on his having botched the job of destroying the evidence (whether or not it was *physically possible* for him to have done otherwise).

Rather, Dummett's two principles amount to something like this:

(3P) For all times s and t, if s is earlier than t then there *could have been* evidence at t as to whether an event E occurred at s, independently of evidence as to whether an action A is performed at t.

(3F) For all times t and u, if u is later than t then there *could have been* evidence at t as to whether an event E will occur at u, independently of evidence as to whether an action A is performed at t.

The truth of (3P), by ordinary standards, seems clear enough. All

evidence of a given past event (the death of the last dinosaur, say) may in fact be lost. But the unhappy event could have been 'recorded' in some way, and the record could have survived. We would then have had the same access to that event as our descendants will have to the Coronation, say, or Nixon's resignation speech.

However, is (3F) false, by ordinary standards? It is not, at least with respect to many of the kinds of events we take ourselves to be able to bring about. Consider the eating of a cake. Suppose I believe that it is in my power to ensure that a given cake is eaten at a certain time. I may well concede that there *could have been* evidence independent of evidence to my own actions as to whether the cake would be consumed at that time. I could have known that someone else, with a stronger claim to the cake than mine (and the power to prevent me eating it), was planning to have it himself. Or I might have known that the cake was under lock and key, and therefore wouldn't be consumed at all.

As for the eating of cakes, so for most events which humans can affect. The one class of exceptions, I think, are those events (such as voluntary suicides) which only one person has the power to bring about. Otherwise, we may always acknowledge that we *could have had* evidence as to whether any event would occur (independently of any evidence as to our own actions); even though, as it is, we claim the ability to bring about that event.

We now have a puzzle. According to Dummett's analysis, the fact that we ordinarily regard (3P) as true underlies our belief that we can't affect the past. If so, then shouldn't the fact that we also regard (3F) as true (for most of types of events we claim the power to bring about), undermine our belief that we can affect the future? Indeed it should, but fortunately for us the analysis is incorrect, in letter, if not in spirit.

As they stand, neither (3P) or (3F) is incompatible with the corresponding versions of (1) and (2) (the 'sufficiency' and 'power' principles). The reason why not is easily shown by attempting to run the causal loop argument against an ordinary claim to be able to affect the future. Thus suppose I claim the ability to ensure the consumption of a particular chocolate cake at 4:00 p.m. (by means of an action of this kind: lift to the mouth, chew, and swallow). I claim both the power to perform this action, and its sufficiency for the desired effect. Yet I concede that I could have independent evidence as to whether the cake would be consumed at 4:00, of the kind described above. CLA then runs: suppose that in such a case you did have such evidence. You

could then be induced to try to consume the cake when and only when the evidence showed that you would not consume it. Assuming the evidence reliable, then either you could not perform the necessary actions (refuting 'power'), or these actions would not result in consumption of the cake (refuting 'sufficiency'). Either way, your claim must be rejected.

It is clear where the argument fails. I claim only the ability to consume the cake in *certain circumstances* (including those which hold at present). True, there could be evidence of the kind described (evidence that the cake is locked away, for example). But the circumstances will then no longer be those in which I claim the ability.

This may all seem a little trivial. Less so, I think, is the issue to which it leads. In the cake-eating case, my claim to affect the future survived the causal loop argument simply because I could legitimately admit that while in fact I was able to consume the cake, I wouldn't be able to do so in the kinds of circumstances which would provide independent evidence of its fate. A past-directed analogue would thus involve the claim that although in fact I had the power to bring about some past event, the existence of evidence as to whether the event had occurred would so alter the prevailing circumstances as to deprive me of this power. Things could have been otherwise; there could have been such evidence. But then I wouldn't have been able to bring about the event in question.

In other words, the past and future versions of CLA require not (3P) and (3F), respectively, but the following stronger principles:[2]

(3P′) For all times s and t, if s is earlier than t then it could have been the case at t both that the circumstances were of type C and that evidence existed as to whether an event E occurred at s, independently of evidence as to whether an action A is performed at t.

(3F′) For all times t and u, if u is later than t, then it could have been the case at t both that the circumstances were of type C, and that evidence existed as to whether an event E will occur at u, independently of evidence as to whether an action A is performed at t.

The coherence of our ordinary claims to affect the future rests on our willingness to reject (3F′). (3F), as we saw, is generally true, by ordinary standards. Only when E is the kind of event which can only occur as a

result of the actions of one particular person, do (3F) and (3F′) converge: both are then false.

In the past case, analogously, admission of a claim to bring about some past event E requires only that the stronger principle (3P′) be held to be false. If the event is of the 'one-agent' kind just mentioned, (3P) will also turn out false; otherwise, (3P) may be held true, just as (3F) generally is. In illustrating the form of a claim to affect the past, Dummett chooses such a one-agent action: a tribal chief (and no one else, it seems assumed) claims to be able to act to ensure the success of a hunt which has taken place some days earlier. Nothing in Dummett's case rests on the one-agent restriction; but perhaps it is this accidental feature of his example which explains his failure to observe that the principle whose falsity enables us to claim to affect the future is (3F′), rather than (3F). This, and the striking fact that in the past case, it is far from obvious that (3P) is not strong enough for CLA. We take it for granted, in effect, that (3P) guarantees (3P′). Why?

3.

The answer, I think, lies in the nature of the processes by means of which we ordinarily obtain information about the past and the future. We are taking it for granted that only in the past case do the physical processes which give us evidence about temporally distant events have what may be called the *irrelevance property*: roughly, the property that the process concerned not only provides evidence as to how the object event or system *is*, but also as to how it *would have been*, had the evidence-yielding process not been present.

I shall call a process which yields evidence as to the occurrence of a temporally distant event a *determination*; I shall call a determination an *inspection*, if it does have the irrelevance property, and an *intervention*, if it doesn't. (Clearly this notation has some basis in ordinary usage.)

More formally, a determination D of a system S with respect to its possession of a property P at a time t, has the irrelevance property if and only if:

(4) There is some $S(P, t)$ which is both the state of S with respect to P at t revealed by D, and what the state of S with respect to P at t *would have been*, had D not been performed on S.

Perhaps the most important characteristic of inspections is that the information they provide about systems or events of a certain type may be generalised to 'unobserved' systems of the same type (subject, at least, to the usual constraints on induction). Interventions lack this feature: if we can't say that S would have had property P at t, even if the determination D had not been made, then we can't conclude that similar but 'undetermined' systems have this property.

A simple example will illustrate how natural it is to accept (4) for past-directed determinations and yet reject it for future ones, even when dealing with what are in themselves very similar pieces of information. We stop a small sample of drivers using a particular road. On arrival, 60% of our sample are wearing seatbelts; on departure, 100% are doing so. We would ordinarily conclude that about 60% of *all* approaching drivers are wearing seatbelts; and of any particular sampled seatbelt wearer (say), that he or she would have been wearing a seatbelt, even if not stopped. We would not conclude that after passing our point, 100% of all drivers are seatbelted; nor that had a given individual not been stopped, he or she would have necessarily have been wearing a belt after passing this spot. In the arrival but not the departure case, our evidence results from a process possessing the irrelevance property.

The inference we make in the arrival case here would not be justified if we knew that drivers were aware that they were going to be sampled, and belted (or unbelted) accordingly. Many actual past determinations lack irrelevance for reasons like this. Nixon presumably wouldn't have delivered his resignation speech had the television cameras not been present, for example. But we take it for granted that we could have evidence of the occurrence of this event *via* a process which did have the irrelevance property. (Indeed there probably is such evidence: perhaps the testimony of a minor technician, without whom Nixon would still have gone ahead.) In many scientific cases the presence of measurement apparatus is held to affect the property being measured; but again it is taken for granted that this effect can in principle be reduced below any given level of significance, and can in any case often be 'corrected for', so as to yield a result to which the irrelevance property does apply.[3]

With this qualification, we take for granted that past but not future determinations are inspections. It is in virtue of this that although (3F) is easily seen not to support the future version of CLA, (3P) does appear

adequate for the past version. Underlying this appearance, I think, is the fact that given claims (1) and (2), (3P) and the assumption that the information processes in virtue of which (3P) is true are inspections together entail (3P′) (which, as we saw, is what CLA needs). Briefly, this is because if the presence of a determination doesn't make a difference to whether an event E occurs, then it can't make a difference to whether the prevailing circumstances are those in which the action A is sufficient for E. The remainder of this section fills in the details.

To use the notation of CLA, let us suppose that George accepts (3P), and that the determinations in virtue of which (3P) is true are inspections. But suppose he rejects (3P′), saying of a situation which is of type C at a time t that although there could have been information available at t (by means of an inspection) as to whether an event E has occurred at an earlier time s, if there had been such information, the situation would therefore not have been of type C. The non-triviality conditions on C imply that George must admit the possibility of situations such that: (i) an inspection has been made at t as to whether an event E has occurred at the relevant earlier time s, and in which this inspection has revealed that such an event has not occurred; and (ii) had it not been for the inspection, the situation would have been type C. Call this a situation of type C'. In such a situation we could induce George to attempt to perform an action A at t.

Suppose first that he were able to do as we ask. Because he accepts that the determination of whether E has occurred at s which has been made in such a situation has the irrelevance property, he agrees that had it not been made, E would nevertheless not have occurred at s. So he is bound to accept that there is a possible situation – i.e., the one we have been describing, but without the determination – in which the circumstances are of type C at a time t, an event of type E does not occur at the appropriate earlier time s, and yet, apparently, George does perform an action of type A at t. In other words, George must accept that he is not justified in claiming (1), because he is bound to acknowledge that the circumstances by type C' we have described would have provided a counter-example to such a claim, had the relevant determination not been made.

It may seem that George can escape this conclusion by claiming that had this determination not been made, he would not have performed an action of type A at t. 'I would only have done A because you asked me to', he will say, 'And you would only have asked me to because you had

made the determination. So if it hadn't been for the determination, I wouldn't have done A.' However, the move succeeds only if it is *certain* that had the determination not been made George would not have done A. This amounts to the admission that in such circumstances the absence of the determination prevents him doing A, which conflicts with his claim that it is in his power to choose one way or the other.

We have not yet dealt with the possibility that when we ask George to perform an action of type A at t in a situation of type C', he is unable to do so. This outcome does not directly refute (2), because (2) makes no claim with respect to situations which are not of type C. But if George holds that the presence of a determination as to whether E has occurred at s is sufficient to deprive him of the power to perform an action of type A at t, then he cannot consistently claim both that (1) and that this determination has the irrelevance property. For in the circumstances we have described, of type C', blaming his failure to perform an action of type A on the existence of the determination will commit him to the proposition that if it hadn't been for the determination, he would have succeeded. From this, and his claim that if it hadn't been for the determination the circumstances would have been of type C, it follows by (1) that if it hadn't been for the determination an event of type E would have occurred at s. Given that we have assumed that the determination has revealed that no such event took place at that time, this means that the determination does not have the irrelevance property. So, as we said, the belief that it does have this property is incompatible for George with the claim that (1).

To summarize, we have shown that someone who accepts (1), (2), (3P), and the assumption that the determinations in virtue of which (3P) is true are inspections, cannot consistently deny (3P'). As we have seen earlier on, CLA shows that a belief that (3P') is incompatible with a claim that both (1) and (2). So although (3P) is not in itself incompatible with (1) and (2), it becomes so in conjunction with the above assumption – which, as we have seen, we ordinarily take for granted.

The situation is exactly parallel for the future case, except that there, of course, we don't make the corresponding assumption (i.e., the assumption that the determinations of future events in virtue of which (3F) is true are inspections, or may in principle be replaced by inspections).

4.

This analysis takes our ability to affect the future to rest on a special kind of ignorance: that ignorance which the availability of inspections of the future would dispel. This might seem puzzling. How can a real *ability* rest on an *inability*? If we had the further ability to inspect the future, shouldn't we be more powerful, not less so?

The answer, I think, is that the further ability would require the world to be different, in a way which would deprive us of one of the two existing abilities which together justify our claim to affect the future. Roughly, a process yielding knowledge at a time s of an event E at u, guarantees that if E occurs at u then B is present at s, where B is the belief that E occurs at u. Now suppose that a free action A at t (after s) is claimed to be sufficient to bring about E at u (so that if A is performed, then E occurs). This entails that if A is performed at t, then B was present at s. But if an action cannot take place without some particular preceding belief, that action cannot be free. In the deliberations of a free agent, it is always possible that other beliefs and desires will play a part in determining whether or not A should be performed. CLA trades on this fact, in effect arranging that such additional beliefs and desires are such as to de-correlate B and A.

In other words, it is a mathematical fact about the world that it can't be such both that a free action is sufficient for an event, and that an agent can find out (by means of an inspection) whether the event in question actually occurs, before deciding whether to perform the relevant action. To imagine a world in which our ignorance of the future is dispelled is to think of a world without free action, or of a world in which actions have no future consequences, or of a contradiction.

5.

The assumption that the past is accessible by means of inspections is so embedded in our thinking that it may be difficult to imagine how things would be if it failed. Perhaps an imaginary example will help. A gambling machine drops a ball into one of two closed boxes. A player bets on which box contains the ball, and may choose which one to open first. (The construction of the machine ensures that the boxes cannot be opened simultaneously.) It turns out that whichever box is opened, it

always contains the ball. Once players realise that this is so, they can clean up (whatever the odds) by always opening the box on which they have bet.

It would seem natural for players to speak of choosing to open box X, *in order to ensure* that the ball *had* dropped into it. Letting 'E' be 'The ball drops into box X' and A be the action of opening box X, players will be justified in claiming (1) and (2). Yet they may acknowledge (3P). For any player could have had evidence that the ball was in X, without knowing whether he or she would open that box: namely, if someone else had already opened X (and thereby ensured that it contained the ball). However players will reject (3P′); they can only ensure E when X has not already been opened.

One might object that it would be possible to locate the ball without opening either box; by X-ray, for example. By CLA, this would defeat a player's claimed ability to affect this state of affairs. However, this objection misses the point. The example is intended simply to illustrate the features required of a physical situation, if a claim to affect the past is to be justified. Obviously we can't hope to conjure up an actual case of backward influence out of such familiar physical components. A plastic model of a brain cannot be expected to think. But the failings of the model do not show the impossibility of what it represents. To show this requires objections of principle; in this case, objections which demonstrate that we cannot reasonably abandon the assumption that the past is accessible by means of inspections.

6.

In his [1982] D. H. Mellor sets out to show that CLA, as propounded by Dummett, does indeed demonstrate the impossibility of backwards causation. He tries to provide the reasons, where Dummett could not, why we are bound to assume that all past events are accessible in the way that CLA requires. In brief, his argument runs like this: it is a *sine qua non* of an event being said to happen at a given spacetime point that it should have effects in the immediate neighbourhood of that point. If there are no effects, then we have no reason to say that the event exists. These effects must for the same reason have their own effects, and so on. On, in fact, to any future observer, for whom the chain of effects provides evidence of the occurrence of the original event.

Mellor's existence criterion is an attempt to formulate the 'no redundancy' condition which undoubtedly constrains scientific theorising. Science is properly suspicious of theoretical entities which appear to be 'free cogs', with no bearing on the phenomena which the theory is seeking to explain. However, I doubt whether Mellor has got the principle right. For one thing, Mellor's criterion seems to depend on a use of Occam's Razor which cuts too deeply for its own good. The same general reasoning would seem to entail that if the supposed contents of a given region of spacetime have no effects outside that region, our ontological corpus is better off without them, and we should apply the Razor. (Indeed, this would apply to the piece of spacetime itself, if its existence didn't have effects elsewhere.) However, elective surgery on this scale threatens bits of the corpus we would rather keep.

To take an extreme example, consider the region of spacetime bounded pastward by the future light-cone of the place and time at which you read this; i.e., the region comprising your relativistic future. Unless there is backward causation, the contents of this region have no effects elsewhere. Yet it would be odd to deny it existence on these grounds. Or, consider what takes place inside the event horizon of a black hole. None of this, except its total mass and charge, has effects outside the event horizon. But this is not ordinarily taken to mean that nothing happens in there (except that a mass and charge 'exist'). Finally, consider some past event or state of affairs, whose effects have by now petered out: the number of rodents in the White House at the time of Nixon's resignation, say. It would be odd to deny existence to such a state of affairs, simply because its effects haven't extended this far.

To all these cases, Mellor might seem able to reply that the criterion for the spatiotemporal location of an event is a local one. As long as an individual event has immediate effects at some spacetime point, it can be said to exist there; no matter if elsewhere (and *when*), these effects are not apparent. This reply would depend on a careful account of the nature of an event, so as to exclude pseudo-events such as 'the formation and subsequent evolution of a black hole', 'the evolution of the universe after 1984' or 'the evolution of the effects of rodents in the White House on the day of Nixon's resignation', with respect to which the initial problem arose. Moreover, it would require an explanation as to why the Occamist principle invoked should apply to only these 'real' events, and not to larger pseudo-events and state of affairs in spacetime. I doubt whether any plausible such refinements will do the

trick. This aside, a purely local criterion will not guarantee the essential connection with our *own* experience. Imagine for example that someone has claimed that there exist universes entirely separate from our own. The contents of these universes may well be acknowledged to have local effects. Yet because they have no connection with events in our universe, an adequate 'no redundancy' condition should disallow them. Local effects may be a *necessary* condition of a causal connection to our experience, but they are far from sufficient.

Mellor appears to be concerned with individual event-tokens, rather than event-types. With respect to purely local effects, it is at least arguable that ordinary events satisfy his existence criterion: perhaps every single event does have (and have to have) local effects. But as the White House rodent case shows, it is far more doubtful whether event-tokens need have *distant* effects. Cannot all trace of an event be lost? Mellor seems to think not,[4] but as I pointed out in section 2, to rest the impossibility of backward causation on this 'archivalist' view is to make such causation far more plausible than we ordinarily assume it to be. For except perhaps in the grip of Laplacian determinism, we find nothing absurd or surprising in the claim that some past events have no effects on present experience. At best we believe that they *could have had* effects; but this is (or is based on) a belief about event-*types*.

It might seem that a criterion in terms of event-types will serve Mellor's purposes just as well. After all, what CLA requires is (3P'), which is in the same sense a principle about event-types (being counterfactual). However, it seems to me that although such a criterion would certainly deal with cases of the White House rodent kind (in which no evidence actually exists, though it might have done), the future and black hole cases are more difficult. They suggest, as is independently plausible, that for many events existence claims rely on more than knowledge of effects: perhaps on knowledge of causes, and perhaps on some sort of theoretical coherence.

For these reasons, it seems to me that the proper 'no redundancy' condition on scientific theories cannot be as stringent as Mellor suggests. This aside, I think that even Mellor's criterion fails to guarantee what CLA requires. For I may admit that an event E has effects where and when it occurs at s, and yet claim the ability to bring it about by an action A at (the later time) t; so long as I deny that it is physically possible for the effects of E to reach me before t. (They may reach me *after* t, of course. Unless they do so, I may have no particular

reason for trying to bring about E, and no direct way of knowing whether I have succeeded.) Thus in the gambling machine case I might say that the event I cause by opening box X – namely the earlier dropping of the ball into box X – has the effect that the ball is in X when I open it. Yet since it is physically impossible for the fall of the ball to have outside effects before one or other box is opened, the chain of effects doesn't provide the information required by CLA.

The limiting velocity of ordinary causal signals which is a consequence of special relativity provides a partial illustration. Suppose I claimed to be able to cause a solar flare to have just occurred, in the last couple of minutes (relative to a specified inertial frame) by means of some magical chant. If solar flares affected the weather here on Earth, such an ability might be very useful. In any case it is certainly possible to determine whether a solar flare has taken place, by means of its effects on the usual instruments. Yet because there is a finite limit to the speed of propagation of these effects, CLA cannot be invoked to show my claim mistaken. It is beside the point here that the claimed influence of the magical chant conflicts with the same relativistic principle. The example is intended merely to illustrate that physical limitations on the propagation of effects can frustrate the access to past events which CLA requires, and Mellor's local effects principle claims to guarantee. (More relevant is the fact that relativity also shows that the claimed effect is not absolutely earlier than the supposed cause. Some reference frames will see it differently. An example in which the effect was absolutely earlier than the cause would therefore require some other physical constraint on the propagation of effects.)

7.

Hence it seems to me that Mellor's existence criterion is powerless to support the use of CLA against a suitably constructed claim to affect the past. This means that we can sensibly ask the following: what happens if we try to apply the criterion to the interpretation of the phenomena of the gambling machine? Or more generally: how should science interpret a case in which a free action appears to guarantee some earlier event or state of affairs? If CLA were successful, these questions would deserve no answer. CLA aims to show that there can be no reliable correlation between free actions and earlier events; that is, no correlation which could justify performing the action in order to

ensure the event. However, we have seen that Mellor's existence criterion does not provide the foundation that CLA requires. The interpretation problem exemplified by the gambling machine remains legitimate, for we have as yet no reason to be certain that we shall not encounter such physical phenomena.

It seems to me that such phenomena will always admit at least two types of interpretation, in addition to the one which invokes backward influence. In the gambling machine case, the first (the *discontinuity* option) will claim that the opening of box X (say) at t does not ensure that the machine dropped the ball into X at s. Instead it causes the ball to move instantaneously to box X (if it wasn't there already). We thus don't influence the fall of the ball, but only its location at the instant of opening. The second (the *indeterminacy* option) will again say that opening X has an instantaneous effect. But rather than moving the ball (if necessary) to box X, its effect is to give the ball a definite location, when up to t it didn't have one. The location of the ball is thus indeterminate before t, becoming determinate only when a box is opened.

Nothing in the phenomena appears to rule out these interpretations. Both would save us admitting backward influence, though at the cost of either instantaneous actions at a distance (and discontinuous change), or indeterminate physical properties. The choice between the three possibilities seems in one sense to be of a kind quite common in science. The theories of physics, in particular, appear to admit a variety of radically different interpretations. Consider, for example, a universe of Newtonian particles, all of equal mass, exchanging momentum in discrete collisions. Such a universe may be re-described without reference to enduring particles. To each point of spacetime at which (on the usual view) a collision occurs, the re-description assigns a quadruple of vectors corresponding to the arrival and departure momenta of (again as the usual view has it) the two particles involved. (If the particles were not of equal mass, the re-description would require a further pair of scalar values, corresponding to the masses of the particles involved.) In such a world no experiment could demonstrate that particles did exist between collisions. Hence nothing in the observed phenomena rules out the non-standard interpretation.

Such cases seem common. It is difficult to say what ordinarily decides them. However, it does not seem necessarily to be any kind of 'no redundancy' condition. If anything, that should pick out those inter-

pretations with the least ontological commitments: in our case, the indeterminancy option, but in the Newtonian case the 'collisions only' re-description. Whatever the actual condition that science employs, clearly it doesn't support such a policy of compulsory instrumentalism.

The Newtonian case illustrates the attraction of continuous, determinate descriptions of the world. It is difficult to say whether this preference is more than a prejudice, perhaps resulting from our first conceptions of independent physical objects. Prejudice or not, it would seem foolish to abandon these features unnecessarily. If indeterminacy or discontinuity is to be preferred to backward influence, we should want to know the reason why. Here I want to dispatch two fallacious 'reasons why', which I mentioned earlier, and which seem to have unwarranted influence in the debate on the interpretation of quantum mechanics.

One is the feeling that backward influence would lead to causal loops, and hence to paradox. The above analysis shows that this is not so, providing the past effects are not accessible by irrelevant determinations before their cause takes place (as in the gambling machine example they are not).

The other is the impression that backward influence conflicts with free will. If it is already the case that the ball is in box X, and it is in whichever box I am going to open, then how can I still be free to *choose* which box to open?[5]

In the absence of irrelevant determinations of the position of the ball, this argument is exactly parallel to a familiar philosophical argument that there is no free will: by the logical principle of excluded middle, all unambiguous statements, and hence in particular statements about future choices, are either true or false; but if it is already true (say) that I will wear my yellow socks tomorrow, then how can it still be up to me to freely choose to do so?

This is sometimes known as the logical determinist argument. Few philosophers have accepted its conclusion. Of the majority, some, it is true, have thought that free will is only to be saved by denying the argument's main premiss: in particular, by denying that statements about the future have truth values. However, the most popular view is that the argument is somehow mistaken; that free will is quite compatible with general application of excluded middle. To me it seems that the best argument for this view (at least if truth values are to be regarded as temporally-located properties) is along the lines of the

present analysis. The present truth values of statements about our future free choices cannot be determined *in an irrelevant manner* before those choices are made. There is therefore no obstacle to saying that these choices affect the past, to the extent of bringing it about that the relevant truth values are as they are.

Be that as it may, in the absence of inspections of the past event affected the view that backward physical influence would conflict with free will stands or falls with the logical determinist argument. Most philosophers do not find that argument convincing – and nor, I suspect, would most physicists. For example, most physicists (and many philosophers) would agree that such facts as our own future movements in space are correlated with determinate properties of the spacetime manifold. If I'll be in London next Tuesday, then there is a determinate fact about the curvature of spacetime which is associated with my leaving here and getting there. If it is already true that spacetime has this form, then it's already true that I'll go. But few take this to mean I'm not free to choose.

Thus it seems to me that these two philosophical intuitions about backward causation – that it inevitably leads to paradox, and that it conflicts with free will – are without foundation. Unless there is some different objection, the reasonable course would seem to be to save determinateness and continuity, treating the consequent admission of backward causation as a striking discovery about the world.

This brings us to the case which gives immediate relevance to the present discussion: that of quantum mechanics. Here we shall find striking parallels between popular interpretations of the strange consequences of that real theory, on the one hand, and the imagined interpretations of the gambling machine fantasy on the other.

8.

The possibility that irrelevance fails for past-directed determinations in quantum mechanics stems from the feature of the theory known as *superposition*: the fact that if A and B are permissible wave functions for a quantum mechanical system, then so is $c.A + d.B$, where c and d are arbitrary complex constants. The relevant consequences of this theoretical feature are described in the following passage from Hilary Putnam's well-known (to philosophers) discussion of the conceptual problems of quantum mechanics ([1979], p. 138):

To illustrate the rather astonishing physical effects that can be obtained from the superposition of states, let us construct an idealized situation. Let S be a system consisting of a large number of atoms. Let R and T be properties of these atoms which are incompatible. Let A and B be states in which the following statements are true according to both classical and quantum mechanics:

(i) When S is in state A, 100 per cent of the atoms have property R.
(ii) When S is in state B, 100 per cent of the atoms have property T –

and we shall suppose that suitable experiments have been performed, and (i) and (ii) found to be correct experimentally. Let us suppose there is a state C that is a 'linear combination' of A and B, and that can somehow be prepared. Then classical physics will not predict anything about C (since C will, in general, not correspond to any state that is recognised by classical physics), but quantum mechanics can be used to tell us what to expect of this system. And what quantum mechanics will tell us may be very strange. For instance we might get

(iii) When S is in state C, 60 per cent of the atoms have property R, and also get
(iv) When S is in state C, 60 per cent of the atoms have property T –

and these predictions might be borne out by experiment. But how can this be? The answer is that, just as it turns out to be impossible to measure *both* the position and the momentum of the same particle at the same time, so it turns out to be impossible to test *both* statement (iii) *and* statement (iv) experimentally in the case of the same system S. Given a system S that has been prepared in the state C, we can perform an experiment that checks (iii). But then it is physically impossible to check (iv). And similarly, we can check statement (iv), but then we must disturb the system in such a way that there is then no way to check statement (iii).

Putnam introduces what he calls 'The Principle of No Disturbance (ND)':

> The measurement does not disturb the observable measured – i.e. the observable has almost the same value an instant before the measurement as it does at the moment the measurement is taken.

He claims that 'this assumption is incompatible with quantum mechanics'. 'Applied to statements (iii) and (iv) above, the incompatibility is obvious' ([1979], pp. 138–139).

Putnam's principle ND is not the assumption that quantum-mechanical measurements are inspections. The gambling machine case illustrates the difference: if we open box X, and find the ball inside it, then on the backwards influence interpretation we know that the ball was in box X an instant before we opened it. ND holds, though irrelevance fails. (We shall see below that if ND fails then so does the irrelevance principle, though not necessarily a weaker version, which is then significant.) In claiming that ND is incompatible with quantum

mechanics, Putnam has thus overlooked the possibility of a backwards influence interpretation.

Quantum mechanics actually predicts not (iii) and (iv) but the following weaker propositions:

(iii') When S is in state C *and an* R*-measurement is made*, 60% of the atoms have property R; and

(iv') When S is in state C *and a* T*-measurement is made*, 60% of the atoms have property T.

This being so, ND leads to a contradiction with quantum mechanics only if we can assume that the proportion of atoms in a certain state which have property R at a given time, is the same for atoms on which a R-measurement is made at (or immediately after) that time as it is in general (and the same for T). However, this amounts to the assumption that the sorts of measurements involved have the irrelevance property.

As the fact that Putnam is able to take the irrelevance property for granted indicates, interpretations of the consequences of superposition based explicitly on this property's rejection have been discussed very little.[6] This is despite the fact that perhaps the most popular position, the so-called *Copenhagen interpretation*, itself involves a kind of rejection of irrelevance. This interpretation comes in various versions,[7] but its key component is the proposal that quantum mechanical systems simply do not have determinate properties such as position and momentum, except when a measurement is made – and then only in the respect being measured (and perhaps certain functionally related respects). Thus this is the quantum-mechanical analogue of the indeterminacy interpretation of the gambling machine. Note that any such interpretation in fact involves the rejection of irrelevance. It is not that had a given determination not been made, the object system in question might have had some *different* value of the property in question, but that it would not then have had *any* value of this property.

The Copenhagen interpretation's greatest challenge is to establish and justify a boundary between the indeterminate domain of quantum mechanics and the (apparently) determinate domain of our ordinary experience – the so-called *measurement problem*. I think it is fair to say that no satisfactory solution to this problem has yet been given.[8] Note that it results directly from the admission of indeterminateness – no

such problem exists if irrelevance itself is given up, in a determinate framework.

Some of the alternatives to the Copenhagen interpretation parallel the discontinuity option in the gambling machine case. Interpretations of quantum mechanics which admit determinate values of observables (such as position and momentum) are often referred to as *hidden variable* interpretations. We have seen that such theories cannot retain both the irrelevance property and the principle ND, in the face of the consequences of superposition. Advocates of such theories have been as little inclined as proponents of the Copenhagen interpretation to reject irrelevance deliberately, and so have given up ND. This amounts to giving up continuity: a measurement is said to give rise to an instantaneous and discontinuous change in the possessed value of the property being measured. As in our gambling machine case, it is important that this change be such that what the measurement reveals is the possessed value which results from it, rather than the value which immediately precedes it. Various special forces and potentials are introduced to account for such changes. Since these introductions serve no other purpose, and have what are regarded as rather implausible properties,[9] such interpretations have not been popular.

In any case, as I said above, one cannot give up ND without also rejecting the irrelevance property. If the performance of a measurement at a time t produces a value of the property concerned which is different from its value an instant before t, then had the measurement not been made, no such change would have occurred, and the value at t would have been very close to what it is in fact an instant before. (We are assuming here that measurements reveal possessed values, but denying this would seem to admit *any* interpretation.)

However, the discontinuity option does preserve a weakened but significant version of the irrelevance property for quantum mechanical measurements (as it does for past-directed determinations in the gambling machine case): the principle that had a given measurement not been made, the value of the property in question *just before* the time of the measurement would have been what it is in fact. In other words, it restricts the highly non-classical occurrences which need to be explained to the time of measurement itself. No doubt this partly explains why this type of interpretation of quantum mechanics has received more attention than ones based on the outright rejection of irrelevance, even in this weakened sense, and the retention of ND

(given which our two senses of the irrelevance property coincide, in fact). Instantaneously affecting the (distant) present is still considered somewhat more plausible than affecting the past. I can see no sound basis for this preference. In practice it may stem from the feeling that if we give up this weak version of irrelevance we shall be admitting that measurements affect the past properties of systems to which they are applied – a correct feeling – but that this is impossible, for the kind of reasons encapsulated in CLA. If so, then in practice the preference rests on question-begging, for we have seen that CLA requires irrelevance.

On the other hand, there would seem to be a good physical reason for the opposite preference. As mentioned above, discontinuity interpretations of quantum mechanics are difficult to reconcile with special relativity. The instantaneous 'action at a distance' required to account for EPR phenomena appears not to be Lorentz invariant. (Indeed, it follows from special relativity that from some points of view this 'action at a distance' will appear to be action on the past.) A backward influence approach would seem to give rise to no such conflict. Past effects will be confined to the past light-cone, as future effects are to the future one.

An interpretation of quantum mechanics based on the rejection of irrelevance for quantum mechanical measurements and the retention of ND, would parallel backward influence interpretation of the gambling machine phenomena. It would be a hidden variable interpretation, in that it would admit determinate values at all times for all observables. However, it would certainly lack a classical feature much sought after by hidden variable theorists, in that it would not provide individual state descriptions for quantum mechanical systems from which could be derived the result of any *possible* measurement in the immediate future. If 'possible' here is taken in the sense it has in the classical case, then for a given system there will be many possible next measurements, other than the one (if any) which is actually going to be performed. These possible but non-actual measurements may actually be incompatible with the system's individual state description, just as in the gambling machine case the possibility that box Y will be opened may be incompatible with the ball's having actually dropped into box X. This feature gives rise to no contradiction via an argument of the form of CLA, and no conflict with free will, because in so far as the state description does have this kind of consequence, it is not epistemolo-

gically available – until the actual next measurement has been made, by which time it is too late for a problem to arise. (The lack of this classical feature enables such an interpretation to avoid the so-called 'no hidden variable theorems', of which the most powerful are that of Bell [1964] and that of Kochen and Specker [1967]. It is not difficult to show that this is the case, though I shall not do so here. For the case of Bell's Theorem, it is explicitly acknowledged, for example, by Herbert and Karush [1978], p. 314.)

The phenomena described by the propositions (iii') and (iv') above, bear a striking resemblance to those of our gambling machine case. Thus let R correspond to the ball's being in box X, T to its being in Y, and R-measurements and T-measurements to opening X and Y, respectively. Then except that they mention a 60% rather than 100% correlation between measurements and results, (iii') and (iv') exactly describe the key features of the gambling machine. That example was constructed to illustrate the kind of phenomena which would best justify a claim to affect the past. Hence it seems to me that we cannot hope for clearer evidence of backward influence than quantum mechanics gives us, unless exhibited in universal rather than merely probabilistic correlations. The gambling machine demonstrates that even with universal correlations, the evidence will not be free of other interpretations. We cannot hope to simply *observe* backward influence, for it is possible only where 'pure' observation is not; that is, where the irrelevance property breaks down. This means that the case for a backwards influence interpretation of quantum mechanics (as for any physical theory) is bound to rest on the interpretation's theoretical and conceptual advantages: for example (in the quantum mechanics case), its apparent compatibility with special relativity, and its ability to preserve determinateness and continuity. (The fact that the issue will have to be settled in this way does not make it illegitimate, or even exceptional. As I mentioned earlier, many physical theories admit 'non-standard' interpretations which are ordinarily rejected – or not considered at all – for just this sort of reason. The unusual feature of this case is rather that all the options appear to involve giving up some previously cherished conceptual principle. There is no 'natural' choice.)

The above advantages aside, it seems to me that the backwards influence approach promises an elegance and simplicity its rivals conspicuously lack. This would not commend it in the face of sound objections of principle, of course; but if there are such objections, I

have yet to see them. In their absence, the interpretation seems to deserve a better run than it has yet been given.

NOTES

* I am grateful to Jeremy Butterfield, Richard Healey, Frank Jackson, Hugh Mellor, Graham Nerlich, Christie Slade, Jack Smart, and participants in seminars in Cambridge and Canberra, for many helpful comments on earlier versions of this paper.

[1] In Dummett's version the principle is actually that one can know of the past event independently of one's future *intentions*. But if intentions are a reliable guide to future actions this is equivalent to the principle given here; and otherwise, the present version is what the argument requires.

[2] Here we place the same restrictions on C, to avoid triviality, as we did for (1) and (2).

[3] Except perhaps in quantum mechanics; see section 8.

[4] See his [1982], p. 175.

[5] A notable proponent of this argument is J. S. Bell. I heard him use it at a conference in Canberra in 1982, and there is a strong hint of it in his [1981], p. 57.

[6] Work which might be taken in this sense includes that of O. Costa de Beauregard, in a series of papers, including [1976]; D. W. Sciama's short paper [1958]; and C. W. Rietdijk's [1978].

[7] See Jammer [1974], Chs. 4, 6.

[8] See Jammer [1974], Ch. 11, and Putnam [1979], pp. 147–156.

[9] Particularly its apparent incompatibility with the Lorentz invariance requirement of special relativity (see, e.g., Bell [1981], pp. 57–8). See also Jammer [1974], Ch. 7; and Putnam [1979], pp. 140, 145.

BIBLIOGRAPHY

Bell, J. S.: 1964, 'On the Einstein–Podolsky–Rosen Paradox', *Physics* **1**, 195–200.

Bell, J. S.: 1981, 'Bertlmann's Socks and the Nature of Reality', *Journal de Physique* (*Colloque C2*) **42**, 41–62.

Costa de Beauregard, O.: 1976, 'Time Symmetry and Interpretation of Quantum Mechanics', *Foundations of Physics*, **6**, 539–559.

Dummett, M.: 1964, 'Bringing about the Past', *Philosophical Review* **73**, 338–359; reprinted in his *Truth and Other Enigmas*, Duckworth, London, 1978, pp. 333–350.

Everett, H.: 1957, '"Relative State" Formulation of Quantum Mechanics', *Reviews of Modern Physics* **29**, 454–462.

Herbert, N. and Karush, J.: 1978, 'Generalisation of Bell's Theorem', *Foundations of Physics* **8**, 313–317.

Jammer, M.: 1974, *The Philosophy of Quantum Mechanics*, Wiley, New York.

Kochen, S. and Specker, E. P.: 1967, 'The Problem of Hidden Variables in Quantum Mechanics', *Journal of Mathematics and Mechanics* **17**, 59–81.

Mellor, D. H.: 1982, *Real Time*, Cambridge Univ. Press, Cambridge.

Putnam, H.: 1979, 'A Philosopher Looks at Quantum Mechanics', reprinted in his *Mathematics, Matter and Method*, 2nd ed., Cambridge Univ. Press, Cambridge, pp. 130–158.

Reitdijk, C. W.: 1978, 'Proof of a Retroactive Influence', *Foundations of Physics* **8**, 615–628.
Sciama, D. W.: 1958, 'Determinism and the Cosmos', in S. Hook (ed.), *Determinism and Freedom in the Age of Modern Science*, N.Y. Univ. Press, New York, pp. 76–78.

Department of Philosophy
Research School of Social Sciences
The Australian National University
P.O. Box 4
Canberra ACT 2600
Australia