NOÛS 37:3 (2003) 471–497

Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It[1]

KAREN BENNETT
Princeton University, Australian National University

1. The Problem

The basic form of the exclusion problem is by now very, very familiar.[2] Start with the claim that the physical realm is causally complete: every physical thing that happens has a sufficient physical cause. Add in the claim that the mental and the physical are distinct. Toss in some claims about overdetermination, give it a stir, and voilà—suddenly it looks as though the mental never causes anything, at least nothing physical. As it is often put, the physical does all the work, and there is nothing left for the mental to do.

I have purposely left that version neutral between events and properties; slightly different versions arise depending upon whether it is type or token identity that is denied. That distinction will matter a bit later on, but for now I shall continue to speak neutrally in order to bring out the overall shape of the problem.

It is a forceful argument. And the primary reason for this, I think, is that it does not attempt to claim that there is something about the nature of the mental that renders it incapable of causing anything. This means that it is rather different from other worries about the efficacy of the mental, such as those that arise for Cartesian dualists, those that arise in the wake of Davidson's anomalous monism (1969), and those that arise for externalists about mental content (Block 1990). Those worries turn on claims about the failings of the mental—that it is not spatially extended, or is not invoked in the requisite sort of strict laws, or is somehow inappropriately extrinsic. The exclusion problem, in contrast, does not purport to show that mental events and properties are somehow by their nature unsuited to causing anything. It is rather that even if they *are* perfectly suited to causing things, there is nothing around for them to cause.[3]

This can be seen by means of the commonplace observation that the exclusion problem does not get off the ground without the claims about overdetermination to which I made but offhand reference above—in particular, that a) any physical effect that did have a sufficient mental cause would be overdetermined, and b) the physical effects of mental causes are not systematically overdetermined in this way. The reason those claims are crucial is that there is otherwise no reason to think that the mental never causes anything. After all, it is not as though the sufficient physical cause literally drains away the causal power of the putative mental one; the existence of one sufficient cause does not entail that there are no others. The claims about overdetermination are required precisely because the rest of the premises do nothing to establish that the mental has no causal power. So the claim is not that the mental is inherently unsuited for causing anything, but rather that there is a problem even assuming that it *can* cause things. Thus the exclusion problem in a certain sense arises *after* the other problems about mental causation. We can pretend that they have all been solved, that mental events and properties can very well be causes. The question remains: how on earth are we to avoid the claim that they are at best overdetermining causes?

As I see it, then, the issue lying at the heart of the exclusion problem is that there is a tremendous tension between the claim that mental events and properties are causally efficacious, and the claim that they do not overdetermine their effects. The more you go out of your way to establish the full-fledged efficacy of the mental, the more it sounds like its effects are overdetermined. And the more you go out of your way to deny overdetermination—to say that mental and physical causes do not 'causally compete'—the less it sounds like the mental is both genuinely efficacious and genuinely distinct from the physical. This has become an underlying theme of Kim's; I take it that it is behind much of what he says in the second chapter of *Mind in a Physical World* (1998).

This tension is both why the exclusion problem seems so intractable, and why it really needs to be tracted. Now, it can obviously be dissolved by denying one of the primary claims that generate it. We could deny the distinctness of the mental and physical—after all, a lot of people want to use the exclusion problem as an argument for type or token identity, depending on what version of the problem is in question (e.g. Peacocke 1979, 134–139; Schiffer, 1987, 146–154; Kim 1989a, b, 1993a, b). Or we could resign ourselves to denying the efficacy of the mental, in the way that Prior, Pargetter and Jackson use an exclusion problem about dispositions and their causal bases to conclude that dispositions do not in fact do any causal work (1982). And we could in principle deny the completeness of physics, although that is not a particularly popular option.[4]

However, a lot of people want to keep all of those claims. They want to hold fixed completeness, the distinctness of the mental and physical, and the

causal efficacy of the mental, and *still* deny overdetermination. What they want to deny, then, is the claim that lurks in the background—that no effect can have more than one sufficient cause unless it is overdetermined.[5] I shall borrow a label from Terence Horgan (1997) and call this view 'causal compatibilism'.[6] It dates back at least to Goldman's immediate response (1969) to Malcolm's introduction of the exclusion problem to the contemporary philosophical stage (1968), and versions of it have been adopted by quite a lot of people lately (e.g. Blackburn 1991, Burge 1993, Horgan 1997, Mellor 1995, 103–104; Noordhof 1997, Pereboom and Kornblith 1991, and Yablo 1992, 1997). The rest of this paper will be about the prospects for compatibilism—about whether its proponents are right to think that we can have our cake and eat it too. In what follows, then, I am going to assume that physics is causally complete, that mental events sometimes cause physical ones, and that their mental properties are sometimes efficacious in those transactions. I shall also assume that both type and token identity are false (the latter, at least, is contrary to my own inclinations). Given those assumptions, can overdetermination be avoided? Can compatibilists maintain this delicate balance?

I used to think that they could not. I used to think that the exclusion problem provided compelling reason to move back towards identity claims, at least in the token case. However, I am increasingly beginning to think that compatibilism holds serious promise. Yet I also very much think that the burden of proof is on the compatibilist, and that her position needs to be worked out rather carefully. My goal in this paper, then, is to try to develop the compatibilist strategy. It will soon become clear that I have no intention of making life easy for the compatibilist; I am going to make her work for her keep. In the end, though, I will suggest a way to make the strategy viable. I am not entirely convinced by it, but I do think that it is more compelling than anything else that has been suggested.

2. The Compatibilist's Task

The place to start is by pointing out that in order to get anywhere with the question of whether overdetermination can be avoided while preserving both the genuine causal efficacy of the mental and its distinctness from the physical, we have to get a much better grip on what overdetermination *is*. On the whole, this has been rather inadequately addressed in the literature. Those who think compatibilism is a nonstarter—call them 'exclusionists'—tend to simply say that of *course* there would be overdetermination. After all, the compatibilist just said that mental causation always involves two distinct sufficient causes; discussion over. And their unrepentant compatibilist opponents tend to simply say that this is crazy, that these cases do not involve overdetermination at all.

That description of the state of play is obviously a bit of an exaggeration, but sadly it is not too much of one. And, as I have said, the burden is on the compatibilist here. She needs to be able to *argue* that the effects of mental causes are not overdetermined, and to explain *why* they are not. She cannot just announce that they are not, for at least three reasons. First, doing so would amount to simply announcing that the very delicate balance she is after does indeed work out in the end—a stance rather like simply shrugging and insisting that of course we have free will. Second, a simple appeal to intuition—"they don't *look* overdetermined"—will not do, because the relevant intuitions are likely to be tainted. It seems to me that the primary source of the intuition that the effects of mental causes are not overdetermined is the residual belief that their mental causes and their physical causes are the *same*. But the compatibilist is proceeding under the assumption that they are *not* the same, and she cannot just help herself to the benefits of identity for free. Third, the exclusionist at least has a definition of overdetermination up his sleeve—an effect is overdetermined just in case it has more than one sufficient cause—and, in accepting the various other components of the exclusion problem, the compatibilist has acknowledged that this is precisely what is going on whenever the mental causes anything. So if she is going to deny the exclusionist's definition, she had better have something with which to replace it.

If compatibilism is to stand a chance, then, its proponents need to give us some genuine reason to think that the effects of mental causes do not count as overdetermined. They need to provide us with some sort of test, some way of deciding whether or not an effect is overdetermined. Let me emphasize that it matters not at all whether we call this a test for overdetermination, or a test for the *bad* kind of overdetermination. The compatibilist could in principle accept that the effects of mental causes *are* always overdetermined, just not in a bad way—the overdetermination is perfectly acceptable, unsurprising, and unproblematic. This is just a terminological issue. For the sake of convenience, I shall speak as though the compatibilist wants to deny overdetermination altogether. But however we choose to put it, what the compatibilist needs to say is that the mental/physical case is importantly different from the standard textbook examples of firing squads, houses that are struck by lightning at the same moment that someone tosses a lit cigarette into the draperies, and so forth. The compatibilist needs to *break the analogy* between the two types of case.

3. An Irrelevant Disanalogy, and a More Promising One

Just to clear the air, here is a preliminary attempt at breaking the analogy that will not get her very far. One thing people sometimes say, at least in conversation, is that the cases are not analogous for the simple reason that the 'overdetermination' in the case of the mental would be extremely

widespread and pervasive. It would happen each and every time we move our bodies. In contrast, the textbook cases of firing squads and so forth are, by their very nature, *rare*.[7] So, the story goes, it is a mistake to reason as follows: occasional instances of overdetermination are somewhat strange, and overdetermination all over the place is exponentially stranger. Instead, the very fact that it would be tremendously widespread makes it *not be like* those other cases.

But why should the sheer extent of the overdetermination make it any less troublesome? The only answer I can see is that its pervasiveness would give us reason to think that it is not a coincidence, which is another thing people sometimes point to as the relevant difference (e.g. Block 1990, 159). This is true enough. So the pervasiveness of the alleged overdetermination by the mental would indeed give us at least *prima facie* reason to think that whatever is going on there is somewhat different from what is going on in firing squad cases and the like.

However, the problem with this sort of line is that it is not obvious why this is an interesting difference. The fact that it is not a coincidence does not mean that it is not overdetermination. Imagine, if you will, a world in which everything works just the way we like to think it actually does, with genuine causation and everything, except that unlike our world it also contains a meddling, Malebranchian god. This meddling god steps in and causes every effect, even though they already have causes that are both perfectly sufficient and perfectly mundane. Now, this is a world in which *everything that happens* has two distinct sufficient causes—one divine and one nondivine. The 'double-causing' is as pervasive as can be, and it is no coincidence. Yet it is also still overdetermination. Or so says my intuition, anyway.

Perhaps I am going on too long about the appeal to pervasiveness, especially since I have never seen it defended in print. Nonetheless, I do think it is important to see that although such an appeal may be a component of the solution, it is not the component that is doing the work. What *is* doing the work? Well, keep in mind that nobody thinks that the mental/physical case is any more like the meddling god case than it is like the firing squad case. The difference, the compatibilist will say, is that there is an important tight relation between the mental and physical that just does not hold between the two shootings, or between either of the shootings and the actions of the meddling god.

This is a much more promising strategy. Stephen Yablo has been particularly explicit in making this kind of move central to his version of compatibilism (1992), but a similar thought seems to be behind other gestures towards compatibilism as well (see especially Pereboom and Kornblith 1991, and Mellor 1995, 104). And it is by no means a crazy idea. First, many people have pointed out that events on a causal chain do not overdetermine their effects, even though both are causally sufficient for them (e.g., Goldman 1969, 471–473; Kim 1989a, 252; Mellor 1995, 103–105; and

Yablo 1992, 272 and 1997, 255). It certainly looks as though the reason such causes do not overdetermine their effects is precisely the fact that an important relation holds between them—namely, causal sufficiency. Second, notice that no one thinks that the effects of mental causes are *never* overdetermined. That is, nobody is worried about defusing competition between mental causes and spatially distant or otherwise unrelated physical causes. Suppose that I want to raise my arm, and do so—and that, at the same time, somebody else grabs my arm and lifts it up. *That* is clearly overdetermination of the firing squad variety, and does not bother anyone. The only real cause for concern is that my arm's going up might *also* come out overdetermined by my desire and the physical state of my body that 'realizes' or otherwise underlies it. In short, the physical causes that the compatibilist wants to say do not causally compete with mental causes are precisely those that are somehow tightly related to the mental causes.

Thus it looks as though the compatibilist should indeed make some sort of appeal to a tight relation between the mental and the physical. However, simply announcing that there is some such relation that makes the exclusion problem go away is far from the end of the story. What kind of tight relation? And, much more importantly, why think that its existence defuses the threat of overdetermination? Providing the compatibilist with answers to these questions is the task of the rest of the paper.

Yet making any progress on them definitely requires figuring out that test for overdetermination. More precisely, what the compatibilist needs is a necessary condition on overdetermination—one that certain kinds of closely related causes fail to meet. After all, she is not going to disagree with the exclusionist that overdetermination requires at least two sufficient causes. What she is going to disagree with him about is whether that is *all* that is required. So let us back away from the mental/physical case for a while in order to think about what overdetermination requires.

4. The Test for Overdetermination

Happily, there is a very intuitive and widely accepted[8] requirement available. It goes like this:

> e is overdetermined by c_1 and c_2 only if
> (O1) if c_1 had happened without c_2, e would still have happened: $(c_1 \,\&\, \sim c_2) \,\Box\!\!\rightarrow e$, and
> (O2) if c_2 had happened without c_1, e would still have happened: $(c_2 \,\&\, \sim c_1) \,\Box\!\!\rightarrow e$.[9]

I shall make two quick points about what this test does not say before defending what it does say.

First, the fact that it is a counterfactual test does not commit the compatibilist to a Lewis-style account of causation. I am not trying to say *anything* about the deep truth about causation, partly because I do not know what it is, and partly because the exclusion problem does not appear to rest on any particular account of it. Besides, we are certainly allowed to reason about causation in the absence of any account of its nature—consider what we do when we teach Mill's methods to informal logic classes—and that is all that this test for overdetermination purports to do. So the compatibilist can perfectly well endorse the test without endorsing any substantive view about how causation works.[10]

Second, the test is not intended to provide a sufficient condition on overdetermination. I am not claiming that overdetermination is guaranteed by the truth of the counterfactuals—nor even by the truth of the counterfactuals plus the further uncontroversial requirement that c_1 and c_2 be causally sufficient for e. I am only claiming that the test provides a *necessary* condition on overdetermination. That is all the compatibilist needs.[11]

Why, then, should we think that the truth of the counterfactuals in fact is necessary for overdetermination? The main reason is simply that they capture the reasoning we engage in when we want to distinguish cases of genuine overdetermination from cases of joint causation, or from cases in which one of the putative causes is not really a cause at all. Let c_1 and c_2 be the shots fired by two members of a firing squad, and e be the victim's death. If we needed to decide whether or not the death was overdetermined, we would ask precisely whether these two counterfactuals are true. Would the victim have died if the first gunman had fired without the second? Would he have died if the second gunman had fired without the first? If the answer to both questions is 'no'—if both counterfactuals are false—then the death was not overdetermined, for it was jointly caused by the two gunshots. If only one of the counterfactuals is false, at most one of the gunmen is guilty. So the truth of the counterfactuals does play an important role in our willingness to say that some effect is overdetermined. Indeed, it is hard to see how they could *fail* to be true when the relevant e is overdetermined. Counterexamples to the necessity claim are rather hard to come by, as long as we are careful in judging the truth-values of the counterfactuals.

It is important to see, though, that care *is* required here. Consider a putative counterexample that again uses our handy firing squad. Suppose that the first gunman is quite serious about his work, and would only fail to fire his gun if some terribly traumatic event occurred just before he was to do so—the sudden collapse of a beloved commanding officer, for example. But that kind of event would leave the second gunman shaken up as well, and would throw off her aim. Consequently, it looks as though the victim would *not* have died if the second gunman had fired without the first—the second gunman would have missed. Yet despite the apparent falsity of that

overdetermination counterfactual, the death *is* clearly overdetermined; the victim actually got hit with two bullets. So does it follow that the truth of the counterfactuals is not in fact necessary for overdetermination? No. What follows is that this is a bad way to evaluate them. It involves what Lewis calls *backtracking* (1973a, 1979). That is, the reasoning takes the following form: if c_1 had not happened, that must have been because x happened, and if x had happened, c_2 would have happened in such a way that it would have failed to cause e. Now, backtracking evaluations are not always and everywhere wrong, but they are definitely inappropriate in some contexts, and I hereby claim that this is one of them. To get the proper results from the overdetermination test, you cannot backtrack, looking for the *reason* the one event failed to occur. You just imagine its failure to occur, period. So even in this case, if the first gunman had not fired but the second had, the victim *would* still have died. This case does not constitute a counterexample to the necessity claim.

The upshot thus far, then, is that as long as (O1) and (O2) are evaluated properly, their truth does indeed look necessary for overdetermination—and does so entirely independently of the compatibilist's project. The counterfactual test really is a quite natural one. And, crucially, it does in fact allow the compatibilist to fulfill her agenda; it lets her start to get somewhere with the idea that a tight connection between the causes can help defuse the threat of overdetermination. The important point here is the simple fact that certain kinds of tight connection between c_1 and c_2 can affect the truth-values of (O1) and (O2). Thus if I am right that those counterfactuals are critical to our notion of overdetermination, this allows for something like an explanation of how, in some cases, an effect can have more than one sufficient cause and nonetheless not be overdetermined.

Take, for example, the already noted issue about cases in which c_1 is causally sufficient for c_2, which is in turn causally sufficient for e—cases in which c_1 and c_2 are parts of the same causal chain. Our intuitions are quite univocal that these are not cases of overdetermination, and this generates a difficulty for the typical exclusionist claim that any effect with more than one sufficient cause is overdetermined. One common line is to move to the claim that any effect with more than one *simultaneous* cause is overdetermined (e.g. Goldman 1969, 473); a less common line is to add a complicated caveat to the effect that the two causes cannot be such that either is causally sufficient for the other (hinted at in Dretske and Snyder 1973, 290–291). Given the counterfactual test, however, neither restriction is needed. (O1) will be false in the case at hand—if c_1 had happened without c_2, e would *not* still have occurred. The causal chain 'stops before getting there', as it were.

In fact, there are causal chain cases that the counterfactual test handles better than either of the other restrictions to the exclusion principle.[12] Suppose that there is a causal chain leading from c_1 through c_2 to e, and

that c_1 is also *directly* causally sufficient for e. This case violates both the restriction to simultaneous causes and the restriction to causes that are not themselves causally related to each other. However, it does not violate the counterfactual test; both (O1) and (O2) come out true. And this is all to the good, as it seems to me that if this sort of case is possible, the effect is indeed overdetermined. Contrary to the suggestion of some (Goldman 1969, 471–472, Yablo 1992, 272), a causal relation between c_1 and c_2 does not automatically mean that they do not overdetermine e. It depends on the case.[13]

Now, given that no one but Searle (e.g. 1992) thinks that causation is the relation between the mental and physical,[14] this brief discussion of causal chain cases may not look all that directly relevant to the exclusion problem. Yet it not only provides a nice example of a connection between the causes that can help, but also brings to light the fact that the use of (O1) and (O2) imposes limitations on just what kinds of connection matter. Not just any relation between c_1 and c_2 will be relevant to whether or not e is overdetermined, because not just any relation affects the truth-values of (O1) and (O2). Causal relations will not always help, and there are other kinds of relations that quite generally will not help either. For example, a mere counterfactual connection between c_1 and c_2 makes no difference at all. This is simply because the truth of, say, $\sim c_2 \mathrel{\Box\!\!\rightarrow} \sim c_1$ just is not relevant to the truth of (O1): $(c_1 \,\&\, \sim c_2) \mathrel{\Box\!\!\rightarrow} e$. The two are perfectly compatible. All that follows is that some $\sim c_2$ worlds must be skipped over to reach the ones that are relevant to the evaluation of (O1); the closest worlds in which the antecedent of (O1) is true are not the closest worlds where c_2 fails to happen.[15] So the test predicts that a mere counterfactual connection between the causes is not enough to defuse overdetermination. This is as it should be, since it is easy to concoct clear cases of overdetermination in which the two causes are counterfactually connected. Such counterfactual connections might well be the norm in real firing squad cases—if either gunman were to shoot, the other would as well, and if one did not, the other would not either.

However, these reflections show that the test does open the door to the idea that a *tighter* connection between the two causes *would* help defuse the threat of overdetermination—and would do so in a slightly different way than we have yet seen. If one of the causes *guarantees* the existence of the other, there is no issue about skipping over some worlds to get to one where the antecedent of the relevant overdetermination counterfactual holds. There are no further worlds to skip to. To put the point more formally: if one of the causes necessitates the other, if it is at least metaphysically impossible for the one to occur without the other, then one of the overdetermination counterfactuals will come out *vacuous*. And there is something to be said for the idea that the vacuity of one of them means that the effect is not overdetermined. It is certainly true that if *both* of the counterfactuals

are vacuous, overdetermination is not guaranteed—both will come out vacuously true if c_1 is identical to c_2, but it is a conceptual truth, if anything is, that one event cannot overdetermine another all by itself. (I owe this point to Mills 1996, 107). This reason for thinking that their *truth* is not sufficient for overdetermination suggests that it is really their *nonvacuous* truth that is necessary.

Now, this does not itself show that *both* of the counterfactuals have to be nonvacuous; it does not itself show that the vacuity of only *one* of them means that the effect is not overdetermined. However, there indeed is reason to think this. For one thing, the idea that it is metaphysically necessary that one of the causes occurs whenever the other does gives some content to the often-heard idea that despite not being identical, the mental and physical causes are not exactly *distinct*, either. And it also means that there is a sense in which one of the overdetermination counterfactuals is not quite up for discussion—you cannot quite ask what would happen if the one occurred without the other if it just *can't* occur without the other (see Yablo 1997, 257–258). So although I am not going to provide a full-fledged defense of the claim that overdetermination can be defeated by the vacuity of only one of the counterfactuals, I think it is worthy of consideration. At least for the sake of argument, then, let us suppose that overdetermination requires that *both* counterfactuals be nonvacuously true.

At long last, then, we can see what the compatibilist's options for escaping the exclusion problem are. She can either argue that one of the counterfactuals is false, or that one of them is vacuous. To fix the terminology, I shall henceforth let c_1 be the mental cause, and c_2 the physical cause. That is, the relevant versions of (O1) and (O2) are now as follows:

(O1) if m had happened without p, e would still have happened (m & $\sim p$) $\Box\!\!\rightarrow$ e, and
(O2) if p had happened without m, e would still have happened (p & $\sim m$) $\Box\!\!\rightarrow$ e.

Yet even after all this effort, it is unfortunately by no means obvious that having this test in hand will actually help the compatibilist. It is by no means obvious that she can successfully deny that both (O1) and (O2) are nonvacuously true.

That is, I think she *can* deny that both of them are nonvacuously true. But I do not think that it is *obvious* that—or how—she can. There are definite obstacles to doing so, and we need to understand them in order to properly understand how the compatibilist solution works. In the next two sections, then, I shall explain why it appears as though she cannot claim that either (O1) or (O2) is either false or vacuous. I shall then go on to argue that these preliminary difficulties can be avoided, and that the compatibilist may be able to solve the exclusion problem after all.

5. A First Attempt at Falsity

At first glance, it does not look as though the compatibilist can claim that either of the overdetermination counterfactuals is false. Consider (O2) first. Could it be that if *p* had happened without *m*, the effect would not—or at least might not[16]—have occurred? Or, in the property case, could it be that if the cause had had P and not had M, the effect would not have occurred? The prospects look grim.

The problem is that it is hard to see how to claim that (O2) is false without undermining *p*'s putative causal sufficiency for *e*. Here is the line of thought. If *e* would not occur if *p* occurred without *m*, then it certainly sounds as though *p* needs *m*'s *help* to bring about *e*. But how can *p* need *m*'s help if *p* is causally sufficient for *e*? Now, I do not mean to imply that causal sufficiency means that literally *nothing* else can be necessary for the effect to occur; causal sufficiency is only sufficiency in the circumstances. Consequently, there is nothing wrong with *p*'s needing various background conditions and causal intermediaries in order to bring about *e*.[17] The problem is rather that *m* cannot be one of them, on pain of violating completeness. Thus (O2) had better not be false. Or so it seems, anyway; I will actually come back to this and suggest that maybe it *is* false.

Insofar as it works, though, this basic line of argument had better apply to (O1) as well. The details are a little bit different, of course, given that completeness cannot be used to justify the claim that *p* is neither a background condition for nor causal intermediary by means of which *m* causes *e*.[18] But the difference in the details matters less than the sameness of the outline—it is *m*'s causal sufficiency for *e* that apparently rules out the falsity of (O1), just as it is *p*'s causal sufficiency for *e* that apparently rules out the falsity of (O2). Remember, the game we are playing is to see whether the compatibilist can deny overdetermination while holding fixed the full-fledged causal efficacy of the mental. Consequently, the argument that apparently entails the truth of (O2) had better apply to (O1) as well.

This is very much worth emphasizing, because a number of people make a rather different argument for the truth of (O1)—an argument that does not do justice to the fact that the mental cause is *also* supposed to be causally sufficient for the effect. It goes like this. If event *m* had occurred but event *p* not occurred, some other, closely related physical event *p** would have occurred, and *p** would have been sufficient to bring about *e*. Similarly in the property case—if event *c* had property M, but had not had property P, then M would have been realized by some other physical property P*, and P* would have been efficacious in *c*'s bringing about *e*. (See Lepore and Loewer 1987, 639; Mills 1996, 109; and Pietroski 1994, 358–359. For slightly less explicit versions of the argument, see Blackburn 1991, 246–249; Jackson and Pettit 1990a, 204, and 1990b, 114–115; and Yablo 1992, 278).

Now, this had better not be the compatibilist's only reason for thinking that (O1) is true. It is compatible with the mental event or property being utterly epiphenomenal. Notice, for example, that Jackson and Pettit rely on similar reasoning to argue that properties that are *not* causally efficacious can nonetheless be cited in epistemically helpful 'program explanations' of why some effect occurred (1990). But to make the point fully vivid, consider a version of the firing squad case in which the first gunman has a line of back-up gunmen behind him, waiting to fire if for some reason he does not. By the above pattern of reasoning, it comes out true that if the second gunman had fired without the first, the victim would still have died—even if the second gunman actually missed completely, or was firing blanks.

The problem with this style of argument is that it rests upon a way of evaluating counterfactuals that is at least as problematic as backtracking. It assumes that the closest world in which the antecedent is true is a world in which p (or the first gunman's firing) is *replaced* rather than *deleted*, and there has to be a general presumption against evaluating counterfactuals like that, at least in causal contexts. Doing so makes them come out true too easily, and would certainly wreak havoc upon Lewis' counterfactual analysis of causation (a fact of which he is aware; see 1973a, 211, and 2000, 190). The closest world relevant to the evaluation of a counterfactual with a negative antecedent is not one in which a barely different, or even too relevantly similar, event occurs. There are interesting questions here about how to accommodate this presumption against replacement into the semantics of counterfactuals, but addressing them would take me too far afield.[19] I shall simply say that it is clearly akin to the earlier presumption against backtracking evaluations, and that the two together tell us much about how to determine the truth-values of the overdetermination counterfactuals. When you are supposed to imagine c_1 gone, you imagine it *gone*. You do not worry about how the past would have to be different to make it fail to occur, and you do not worry about what else might occur in its place. You simply snip it away as though you had a metaphysical hole-puncher.[20]

The point of this digression is that the compatibilist needs to be careful about just *why* (O1) apparently has to be true. The reason is that there is precisely the same *prima facie* conflict between its falsity and m's causal sufficiency for e as there is between (O2)'s falsity and p's causal sufficiency for e. The replacement argument has nothing to do with it. Indeed, it is the very fact that c_1 and c_2 are supposed to be causally sufficient for the effect that generates the presumption against replacement in the first place.[21] Insofar as the causal sufficiency of P/p for e entails the truth of (O2) at all, it entails the truth of (O2) when evaluated by deletion—and similarly for (O1).[22] The compatibilist must apply the same reasoning to both counterfactuals. Failure to do so would mean that she has fallen into the trap of granting the mental cause only a derivative efficacy, and has lost the game of trying to keep the mental and physical causes on an equal footing.

And that, of course, is the primary danger involved in being a compatibilist at all.

Thus far, then, it looks like the compatibilist cannot claim that either (O1) or (O2) is false on the appropriate deletion reading. As I have said, I shall come back and reevaluate this claim. But first, let us look at the other option—the claim that one of the counterfactuals is vacuous.

6. A First Attempt at Vacuity

The vacuity option might sound more promising. All the compatibilist would need to do is argue either that it is impossible to have the mental event or property without the physical one, or else that it is impossible to have the physical event or property without the mental one. These claims would render (O1) and (O2) vacuous, respectively. Let me start with the former, which is the version of the strategy that looks less likely to succeed. It presumably looks rather doubtful that anybody would want to claim that (O1) is the vacuous one, either for events or properties. The compatibilist is not going to say that it is impossible for event *m* or property M to occur without *p* (P). After all, the main reason most people refuse to identify mental properties with physical ones is multiple realization—it certainly seems as though mental property M could be possessed in a variety of physical ways. And the main reason a somewhat smaller number of people refuse to identify mental events with physical ones is precisely analogous. Any particular mental event could have occurred in a somewhat different physical way—perhaps involving a few different neurons than it actually does—but the related *physical* event could not. The modal difference ensures their distinctness by Leibniz's Law (see Boyd 1980, 99–100; Kripke 1980, 147–148; Pereboom and Kornblith 1991, 131–132; Yablo 1992, 268–270). So given that these arguments against type and token identity are probably why the compatibilist is forced to be a compatibilist in the first place, it looks unlikely that she will want to insist that it is (O1) that is vacuous.

Admittedly, matters are a bit more complicated than that, given my own stubborn insistence upon deletion readings of the counterfactuals. It might look as though that insistence opens the door to the claim that (O1) is in fact vacuous after all. However exactly the presumption against replacement evaluations is accommodated, what we are supposed to be thinking about is not just the possibility of event *m* or property M's occurring without *p* (P) in particular, but rather the possibility of *m*'s (M's) occurring without anything relevantly *p* (P)-like at all. Consequently, it may look as though (O1) really *is* vacuous when properly evaluated. However, this is not in fact true.

The claim that it is impossible for property M to occur without any relevantly P-like property basically amounts to the claim that physicalism is necessarily true. Yet most people think it is only contingent

(e.g. Lewis 1983, 362; Chalmers 1996, 41–42; Jackson 1998, 11–12). Though there may not be any souls *here*, there are worlds in which there are, and in those worlds things can have M without having any physical properties at all. So property M can indeed be instantiated without any physical property, and my insistence on deletion readings does not render the property version of (O1) vacuous. The reason why it also does not render the token version of (O1) vacuous is somewhat trickier, but the same basic point applies—making such a claim requires denying that physicalism is a merely contingent truth. In the interests of keeping the main thread of argument on track, however, I will relegate those details to a footnote.[23] The point is just that, regardless of these complications about replacement and deletion, it is not plausible to claim that (O1) is vacuous.

But what about the other overdetermination counterfactual, (O2)? Surely it looks more promising to claim that it is impossible for the physical cause to happen without the mental one. Lots of people who deny that the mental and the physical are identical think that some kind of 'upwards' necessitation relation holds between them, whether it be supervenience, determination (Yablo 1992), or something else along those lines.[24] It is quite popular for those who deny type and/or token identity to claim that this kind of asymmetric dependency holds—that the physical necessitates the mental, even though the mental does not return the favor.

Sadly, though, it is not obvious that this is true either. The event case is clearest, so I shall begin there. The problem is that the essences of mental and physical events do not nest in the way that this kind of story demands. Even though the most popular argument against token identity is the multiple realization-ish argument against the *downwards* necessitation of the physical event by the mental one, there definitely are arguments from the other direction. That is, there are arguments against token identity that appeal to properties that mental events have essentially, but their corresponding physical events have only accidentally. And whether or not they are at the end of the day a good idea, the compatibilist is going to have a hard time rejecting them, given that they are in precisely the same style as her own arguments for distinctness. I shall just mention one, which is due to Tyler Burge (1979, 111). Start with the claim that contentful mental states have their contents essentially. A belief that ϕ could not be a belief about something *else*. Add in some version of externalism—what a belief is about depends upon various facts about the outside world. From those two claims, it follows that contentful mental states bear certain relations to the outside world essentially. The problem is that it certainly does not look like the physical events to which they are most closely related bear those relations to the outside world *essentially*. This particular pattern of neural firings could occur in a petri dish.

Let me be very clear that the issue here really does not have much in particular to do with mental content; there are a number of other arguments

that can be made against the upwards necessitation of mental events by physical ones.[25] Burge's argument is but one way of pointing out that physical events are not *less* modally flexible than mental ones. They are just modally *different*. The essences of mental and physical events crosscut each other, which means that *neither* overdetermination counterfactual is vacuous. *p* can occur without *m* just as easily as *m* can occur without *p*.[26]

What about properties? Thus far I have been talking about events and their essences. Basically the same problems arise; (O2) is no more plausibly vacuous in the property case than in the event case.[27] The sorts of physical properties that get invoked in the causal claims that lead to the exclusion problem—and thus that get invoked in the overdetermination counterfactuals—do not quite necessitate mental properties. The instantiation of the property *being a C-fiber firing* does not guarantee the instantiation of the property *being a pain*; again, C-fiber firings can occur in petri dishes. The point is also quite clear when mental properties are thought of in functionalist terms, as second order properties. Something's having the property *being a C-fiber firing* does not guarantee its having the property of having some property or other that plays the pain role, even if C-fiber firings actually *do* play the pain role. A property that actually plays a certain role need not do so. Consider the classic lock and key example about dispositions. It is easy to make my key lose the disposition to unlock my door without altering the physical properties of the key at all—I just have to change my lock. So the instantiation of the first order physical property does not guarantee the instantiation of the second-order dispositional one.

I realize that at this point a fairly obvious objection will be raised. It presumably looks as though I am being awfully stupid about what physical properties are. In particular, it presumably looks as though I am foolishly assuming that all physical properties are intrinsic. Yet it is extremely important to see that I am not assuming that at all. My point here only has to do with the kind of physical property that is invoked in typical versions of the exclusion problem, and hence in typical substitution instances of (O2). The exclusion problem typically goes like this: given that this event's having the property *being a certain pattern of neural activity* was efficacious in bringing about my raising my arm, how could its having the property *being a desire to raise my arm* also be causally efficacious? To defuse *that* competition by declaring the relevant version of (O2) vacuous, we would have to say that it is impossible for anything to have the neural property without also having the desire property. And *that* is just not impossible.

Yet the idea that lies behind the objection is a good one, and the time has come to stop making life difficult for the compatibilist. Now that we have seen both *that* it is not easy for her to claim that the overdetermination counterfactuals are either false or vacuous, and *why* it is not easy for her to do so, we have come far enough to see the way out. In fact, we have come far enough to see *two* ways out—though, as we shall see, they are very

closely related. The first defends the claim that (O2) is vacuous, and the second defends the claim that it is false.

7. Rescuing the Vacuity Claim

The first solution is the one that arises immediately from some of the things I just said against vacuity. Maybe it could be maintained that even if event *p* or property P does not necessitate *m* (or M), there is some other physical event or property that *does*. If we are allowed to be fairly generous about what events there are, surely some event *p** can be found whose essence either includes or necessitates *m*'s essence. Yablo makes a move along these lines (1992, 266–268). And similarly for properties—even though the physical property we initially fixed on does not necessitate the mental one, there presumably is a richer physical property that *does*. Just build on P by conjoining in the laws of nature, and other facts about the outside world. So even though there is certainly *a* fairly complicated physical property of my door key that does not guarantee its having the disposition to unlock my door, there is presumably a *more* complicated one that *does*—namely, a partially extrinsic one that includes certain features of the configuration of my lock.

Yet although this line sounds reasonably promising, it faces at least two problems that must be noted. One is that taking this line threatens to undercut the motivations for insisting that the mental and physical are distinct.[28] This is a problem not because the distinctness claim is something to hold onto at all costs, but rather because it is a crucial part of the compatibilist position—without it, she would not face the exclusion problem, and would not need to be a compatibilist in the first place. So why does it threaten to undercut the motivations for distinctness? Well, to say that we get to concoct—or dig around until we find—physical events or properties that do necessitate the mental ones is, in effect, to say that the physical events or properties invoked in the arguments against upward necessitation of the mental by the physical are the *wrong ones*. And that means that, as arguments against type or token identity, they are *just missing the mark*. For example, it seems to follow that when Burge argues against token identity by arguing that mental events essentially have certain relational properties that their associated physical events have only accidentally, he is simply looking at the wrong physical events. If there are physical events that *do* necessitate mental ones, then those physical events *do* have the relevant relational properties essentially, and *they* are the ones on which the denier of token identity should be focusing. And while there are other arguments against identity—namely, the more popular multiple realization-ish arguments against the downward necessitation from the mental to the physical—the same point presumably applies, *mutatis mutandis*, to them. The upshot is just that if the compatibilist gets to find physical events and

properties with the modal profiles that suit her purposes as far as denying overdetermination goes, it is hard to see why the identity theorist—either type or token—is not allowed to find ones with modal profiles that suit *his* purposes.

I think that this worry is both interesting and serious. However, I also think that it has much less to do with this particular version of the compatibilist solution to the exclusion problem than it does with the claim that mental events and properties are not identical to physical ones. The issue here is a quite general one having to do with what physical events and properties there are, and with the difference between claiming that some mental event or property is not identical to some particular physical event or property, and claiming that it is not identical to *any* physical event or property. This general issue is just brought to the surface by the fact that proponents of this solution must explicitly claim that there are lots of physical events and properties, including ones that modally match the mental in at least one direction. So this worry is not really particular to the current context.

The real problem, the one that *is* particular to the current context, is different. The real problem with rescuing the vacuity claim in this way is that it requires saying that the original physical event or property—the one we initially fixed on, the one that does not necessitate *m—is not in fact causally sufficient for e*. If the compatibilist does *not* say that, she has not said *anything* to moot the overdetermination of *e* by *m* and *p*—she has simply changed the subject, and insisted on the vacuity of a different counterfactual. She has also managed to introduce a new overdetermination worry about *p* and *p**. What the compatibilist has to say, then, is that if a mental cause is efficacious in bringing about some effect, the only physical causes that are also efficacious in bringing about that effect are ones that necessitate the mental cause. That is a fairly hefty thesis, especially since it entails that much of our talk about physical causation is misguided. I shall return to this point shortly, for I am by no means intending to lay the vacuity strategy decidedly to rest. However, let us first look at the other option for saving compatibilism.

8. Rescuing the Falsity Claim

To see the alternative strategy, set aside the idea that the vacuity option can be saved in the manner just sketched. Instead, consider the idea that the very reasons to deny that (O2) is vacuous *undercut all reason to think that it is true*. The compatibilist can argue that although *p could* occur without *m*, it would no longer cause *e* if it did. That is, although there *are* worlds in which *p* occurs without *m*, they are different enough from the actual world that we have little or no reason to think that *e* would occur there.

Here is an extreme example of what I mean. Any world in which my current pattern of neural activity occurs in a petri dish (or a vat) is just *not* a world in which you would expect it to cause whatever it actually causes—my raising my hand, say. So if *that* is the kind of world needed to witness the truth of the antecedent of (O2), the counterfactual as a whole is simply false. The p & $\sim m$ world is just not an e world.[29] Or consider the lock and key example again. There are worlds in which my key has all of the intrinsic physical properties it actually has, but lacks the disposition to open my door—worlds in which either the laws of nature or, less drastically, my lock is changed. Yet in those worlds we would not at all expect the properties of my key to be causally efficacious in the same causal transactions it is here—particularly not in opening my door. That is pretty much analytic, and the point can clearly be generalized to all first order properties that actually realize second order functional ones. In worlds in which they do not realize the relevant second order property, they obviously do not cause the things specified by the causal role by which the second order property is defined. The claim, then, is that the physical cause can indeed happen without the mental one—but that if it did, it would no longer be up for causing e.

This way of arguing that (O2) is typically false does not involve backtracking, or any funny business of that kind. The reasoning does not parallel that of the traumatized gunman firing squad case discussed earlier. The claim is not that if the mental cause had not happened, that would have been because of the occurrence of various *prior* goings on that changed what p could cause. The claim is rather that if the mental cause had not happened, that just *constitutively involves* various changes in the world that change, or at least may well change, what p causes. Thus the suggestion I am making does not violate my earlier insistence that when we are supposed to imagine m gone, we are supposed to imagine it *gone*, just snipped away. I am not relying on replacement or backtracking here. The point is rather that, barring a metaphysical miracle, m's being gone partially just *is* these other changes.

This fact also enables us to see what is wrong with the thought that (O2)'s falsity would undermine p(P)'s causal sufficiency for e. Back in Section 5, I said that if e would fail to occur if p happened without m, then it certainly looks like p needs m's *help* to bring about e. Yet p surely cannot need m's help if it is causally sufficient for e—that would at best mean that m is some kind of background condition or causal intermediary, which violates completeness. As I indicated at the time, however, that argument is not quite right. The falsity of (O2) need *not* be taken to mean that p needs m's help in this sort of way. There is a different way to understand the fact that e would not occur if p occurred without m—a way that does *not* give rise to the idea that p 'needs m's help', and that therefore threatens neither completeness nor p's causal sufficiency for e. The alternative is this: the conditions that must hold for p to bring about e—physical conditions, note—*are basically*

the same as the conditions in which p necessitates m. So if *p were* to occur without *m*, those conditions would not hold—and *p* would not, or at least might not, cause *e*. And that does not mean that *p* does not *actually* cause *e*. (O2) can indeed be false compatibly with *p*'s causal sufficiency for *e*.

9. Tying Them Together

What is the relation between the two strategies that I have offered the compatibilist? They are, at bottom, very similar. Both turn on the idea that there is an interesting connection between the relation that holds between mental events and properties and the physical ones that 'underlie' them, and the relation that holds between the physical cause and the effect. And the fact that one strategy says that (O2) is false while the other says that it is vacuously true is not indicative of any genuine tension between them. The two strategies actually focus on different substitution instances of (O2). Thus the counterfactual claimed to be false—namely, $(p \ \& \sim m) \ \Box\!\!\rightarrow \ e$—is not the same as the counterfactual claimed to be vacuous—namely, $(p^* \ \& \sim m) \ \Box\!\!\rightarrow \ e$. *p* and *p** are different events (or properties).

While *p* is an example of the sort of physical event and property we typically talk about, *p** is not. *p** is a rather complex physical property, or a physical event with a rather extrinsic essence. Events and properties like *p** are not the focus of our everyday talk; they stand to the referents of terms like 'neural firing' as table-shaped objects that are essentially located in one bedroom apartments stand to the referents of the English word 'table'. The vacuity strategy, as I have already noted, changes the subject. The *reason* it changes the subject, of course, is that it is only these kinds of physical events and properties that necessitate the mental ones. *p** is precisely the event or property created by adding to *p* the various other physical occurrences that are needed to guarantee *m*.

Notice, though, that because the vacuity strategy must involve the claim that it is only *p**—not *p*—that is causally sufficient for *e*, it follows that whatever constitutes the difference between *p* and *p** is *also* precisely what *p* would need to bring about *e*. Thus *both* strategies claim, implicitly or explicitly, that the conditions in which the physical event or property occurs with the mental one are the same as the conditions in which the physical event or property manages to bring about the effect. The difference is just that the vacuity strategy packs those conditions into the physical event or property itself, and the falsity strategy does not. The only real difference between the two, then, is the notion of causal sufficiency on which they rely. The vacuity line relies on a rather strict notion, while the falsity line allows us to continue on with our somewhat sloppy notion of causal sufficiency according to which reasonably normal physical events like patterns of neural firings count as causally sufficient for action.

This means that the two responses can be combined to yield a rather powerful answer to the exclusion problem. The compatibilist response is arguably best run as a dilemma about causal sufficiency. Suppose first that we adopt a rather strict notion, according to which the only thing that counts as causally sufficient for an effect is a whole big package, consisting of what we might have intuitively thought of as a cause *plus* all necessary background conditions and causal intermediaries by means of which it brings about its effects. On that kind of notion, only rather complicated and partially extrinsic physical events or properties will ever causally compete with mental ones, and thus will ever be invoked in substitution instances of (O2). And such physical causes will indeed necessitate their mental competitors, and the relevant counterfactuals will come out vacuous. Now suppose that we instead adopt a rather more permissive notion of causal sufficiency—a notion on which an event or property can count as causally sufficient for an effect even though it requires some set of background conditions and intermediaries in order to bring that effect about. On *this* kind of view about causal sufficiency, much more intuitive physical events and properties will be invoked in substitution instances of (O2), and those substitution instances will not in general be vacuous. However, they will typically be *false*, for precisely the reasons rehearsed above. Thus whether causal sufficiency is understood in the strict way or the permissive way, the compatibilist has an answer.

10. Concluding Thoughts and Lingering Worries

There are, of course, some remaining questions and concerns about the compatibilist solution that I have outlined. For example, to what extent do the suggestions I have made apply to (O1) as well as to (O2)? I do not think that they do; much is lost in the translation. I shall not argue this here, however. After all, if I am wrong, so much the better for the compatibilist. But it is worth noting that there is an obvious dialectical advantage involved in claiming, as I have, that the problem lies with (O2) rather than (O1). As I have repeatedly emphasized, the primary danger the compatibilist faces is inadvertently undercutting the assumption that the mental has causal power. She is running a constant risk of sounding as though the mental is really just epiphenomenal. Consequently, if she can claim that it is the *physical* that in a certain sense needs the mental, rather than the other way around, she is in a much better position. And that is precisely what challenging (O2) while leaving (O1) alone amounts to.

But the real question, of course, is whether the compatibilist has genuinely succeeded in escaping the exclusion problem. I shall now consider two objections to her solution. The first of these claims that the compatibilist's challenge to (O2) has not succeeded. After all, the careful reader will surely have noticed that everything I have had to say about how to insist upon

either the vacuity or falseness of (O2) turns on an emphasis on external conditions. In particular, it turns on thinking that the only reason that particular physical events or properties fail to necessitate mental ones is that the mental ones are in various ways more critically *extrinsic*. But it may well be wondered whether that is true. What, you ask, about zombie worlds?

Here, then, is a further lesson about the exclusion problem—the compatibilist has to deny the genuine possibility of zombie worlds. If there is a minimal physical duplicate of our world that is devoid of mentality (or, at least, is devoid of consciousness), then neither of the solutions I have suggested gets off the ground. If it is possible to strip the mental off the world in that way, there simply are no necessitation relations between the physical and the mental. It is just not true that we can find rich and complex physical events or properties that necessitate mental ones, and it is just not true that the conditions in which a physical cause brings about its effects are pretty much the same as those in which it necessitates the mental cause. After all, the possibility of zombie worlds entails that everything (physical) that actually happens could happen just the same if there were no consciousness. However, the completeness of physics does not entail that; it is perfectly compatible with completeness that not everything (physical) *would* happen just the same. That is basically the central insight of the version of compatibilism I have suggested. In order to avail herself of that insight, then, the compatibilist must deny the genuine possibility of zombie worlds.

There is of course a burgeoning literature on how to go about doing that (e.g. Balog 1999, Block and Stalnaker 1999, Loar 1999, Yablo 2000), and I am neither going to rehearse nor defend those arguments here. My only point is that the compatibilist needs to join their camp. But this should come as no surprise. People have known for some time that the view of the mental engendered by belief in zombie worlds is rather unfriendly to mental causation (e.g. Chalmers 1996, 150–155; see also Shoemaker 1975 for related worries). It should not be news, then, that the friend of zombies—the nonreductive *non*physicalist—has things worse than the nonreductive physicalist, who indeed has a shot at a successful compatibilism.

The objection from zombie worlds is an objection to my claims about the status of (O2). The second objection is different. Here, the idea is to *accept* those claims, and instead deny the moral I have drawn. A diehard exclusionist, that is, might commandeer my arguments about the status of (O2) to reject my counterfactual test for overdetermination. The idea would be to say that the primary reason for thinking that the nonvacuous truth of the counterfactuals is necessary for overdetermination is simply that they are nonvacuously true in textbook firing-squad-type cases, and that the reason for *that* is simply that they provide a reasonably good test for causal sufficiency. The exclusionist might claim, that is, that (O1) and (O2) typically come out true in cases of overdetermination *because* c_1 and c_2 are both causally sufficient for e.

Consequently, what he might say is that if I am right that the counterfactuals can be vacuous or false without undermining the sufficiency of the supposed causes, all that follows is that their nonvacuous truth is not in fact necessary for causal sufficiency. And what *that* means is that there is no longer any reason for thinking that their nonvacuous truth is necessary for overdetermination. He could argue, in short, that the proper moral to draw from this paper is not that effects can have more than one sufficient cause without being overdetermined, but rather that the nonvacuous truth of the counterfactuals is not in fact necessary for overdetermination. The compatibilist, then, has not managed to show that the physical effects of mental causes are not overdetermined. She has instead shown that the test she is wielding is not a very good test.

This is a reasonable line of thought, but the compatibilist should nonetheless not be moved by it. The exclusionist is simply digging in his heels and insisting upon the good old-fashioned definition of overdetermination. He can use the word 'overdetermination' in this way if he likes, but all that follows is that the compatibilist will have to coin a new term for the particular kind of overdetermination involved in firing squad-type cases, and rephrase her conclusion as the claim that mental causation does not involve *that*. The compatibilist's task, remember, was to provide a well-motivated and non-handwaving way to break the analogy between the mental/physical case and those other cases. And as long as I am right about the status of (O2), she has indeed accomplished that. She has discharged her burden of proof, and provided a genuine reason for thinking that the tight relation between the mental and the physical makes a difference.

This hardly means that there is no longer any cause for concern about mental causation, of course. I have done nothing to establish that mental events and properties ever really are causally sufficient for anything. I have not shown that the mental actually *can* cause things; I have merely argued that the assumption that it can does not generate massive overdetermination. It is also worth reiterating that I am but a recent convert to compatibilism, and I cannot claim to have entirely assuaged my own lingering exclusionist intuitions. But I do think that the strategy I have suggested is the right one for the compatibilist to adopt. One of its main advantages is the simple fact that it takes seriously the idea that she must *argue* that mental causation need not always be overdeterministic causation, and thus insulates her from the charge that she has simply refused to acknowledge the force of the exclusion problem.

Notes

[1] This paper has gone through a number of incarnations, and I have received a lot of help at various stages. I presented a very distant ancestor at the Michigan candidacy seminar, a

somewhat less distant ancestor at the Metaphysical Mayhem V, and a much more recent version at NYU, my graduate seminar at Princeton, and the ANU. I would like to thank the members of all of those audiences for helpful discussion, particularly Ned Block, Cian Dorr, Kit Fine, Ned Hall, Benj Hellie, David Lewis, Barry Loewer, Gideon Rosen, and Stephen Yablo. I would especially like to thank Sydney Shoemaker, my commentator at the Mayhem, for insightful criticisms that eventually led me to change my mind completely. I used to argue against compatibilism; here I defend it.

[2] Malcolm 1968, Peacocke 1979, Schiffer 1987, and Kim 1989 are among those who have raised the problem; many others have tried to answer it.

[3] As far as I know, Fodor is the only person who has explicitly noted that showing that the mental is *up for* causing things leaves open the possibility that it never in fact *does* (1989, 142).

[4] Note, however, that Sturgeon 1998 has recently given this line an interesting twist. He claims that completeness is only true given a sense of 'physical' that is different from the sense of 'physical' invoked in the claim that mental events and properties sometimes have physical effects. On his view, then, the exclusion argument equivocates.

[5] This is often called 'the exclusion principle'. See, for example, Kim 1989a, 239 and 250–253; and Yablo 1992, 247. The overdetermination clause is often implicit, or relegated to a footnote; not everyone emphasizes that aspect of the exclusion problem as much as they should.

[6] This may not be *quite* what Horgan means by the label, because he does not explicitly worry about overdetermination. His 'causal compatibilism' is the view that "mental causation *via* nonphysical properties can co-exist with physical causation even if the physical realm is causally closed" (1997, 166).

[7] I am obviously assuming that events are not so modally fragile as to render overdetermination impossible.

[8] Mills (1996, 107) uses the test very explicitly. In general, though, the counterfactuals are not put forward as an official test for overdetermination, but are rather mentioned in passing, given as a definition of 'screening off', or implicitly relied upon in explaining the problem overdetermination (or preemption) tends to pose for various theories of causation. To select some citations almost at random: Horgan 1987, 508–509; Kim 1989a, 252 and 253; Lepore and Loewer 1987, 639; Lewis 1973a, 193, and 2000, 183; McDermott 1995, 523–524; Mellor 1995, 101. Notice too that those who think that overdetermination is impossible because they think that events are extremely modally fragile implicitly rely on this sort of test.

[9] This is admittedly more appropriate for the event case than the property case, but it is easily modified. One way to do so would be to use something like

$(O1_p)$ $P_1c \,\&\, \sim P_2c \;\Box\!\!\rightarrow e$
$(O2_p)$ $P_2c \,\&\, \sim P_1c \;\Box\!\!\rightarrow e$,

or we could use a slightly different version that does not assume that the same event has (or is an instantiation of) both properties.

[10] Proponents of nomological accounts of causation should not object to the reliance upon counterfactuals; it might be the case that the counterfactuals hold precisely because certain strict laws do.

[11] It is the *anti*-compatibilist who is in need of a sufficient condition on overdetermination. To argue that there is no possible way for compatibilism to work, he must say that the package of claims that the compatibilist is holding fixed *guarantees* that the effects of mental causes count as overdetermined.

[12] Thanks to David Lewis for making me think about these trickier causal chain cases.

[13] Peacocke appears to make a similar claim (1979, 136).

[14] And notice that the fact that causal relations between c_1 and c_2 *can* help moot overdetermination does not obviously get Searle out of the exclusion problem. His picture is more

like the second causal chain case described above. Further, he in fact faces a version of the overdetermination worry even if he rejects the completeness of physics. See Kim 1995.

[15] *Mutatis mutandis* for the other counterfactual connections that can hold between c_1 and c_2.

[16] Strictly speaking, to claim that (O2) is false is really to claim that if p had occurred without m, e *might* not have occurred. That is, on Lewis' semantics (1973b), the denial of (O2) is $p \,\&\, {\sim}m \diamondsuit\!\!\rightarrow {\sim}e$, not $p \,\&\, {\sim}m \,\Box\!\!\rightarrow {\sim}e$. Since nothing I have to say turns on this, I will ignore it in what follows, and stick to the simpler 'would not' talk.

[17] This is why one of the counterfactuals can perfectly well be false in the causal chain case discussed earlier.

[18] I doubt that much needs to be said to defend the idea that p is neither a background condition for nor causal intermediary by means of which m causes e. It simply does not sound right to say that p is a background condition—whatever the relation between the mental and the physical, it is not the same as that which holds between the striking of a match and the presence of oxygen. And the compatibilist clearly cannot say that p is a causal intermediary by means of which m causes e. Making that move might enable her to say that e is not overdetermined, but only by pushing the problem elsewhere. If p is a causal intermediary, then m must be causally sufficient for it—but p is physical, and, by completeness, must have a sufficient physical cause p_{-1}. So then p has two distinct sufficient causes, m and p_{-1}, and *it* is in danger of counting as overdetermined. The exclusion problem remains untouched.

[19] There are various possible ways of doing this. We could modify the standard semantics for counterfactuals, and say that those with negative antecedents are true just in case the closest *deletion* worlds in which their antecedents are true are ones in which their consequents are true. Or we could keep the standard semantics and instead insist, with an earlier incarnation of Lewis (1973a, 211), that the deletion worlds should typically be counted closer to the actual world than replacement ones. Or we could keep *both* the standard semantics and the perhaps more intuitive standards of similarity, and instead be more careful about what counts as the minimal change from actuality needed to make the antecedent true. The most plausible way to do this is to say, with Lewis (2000, 190), that what the counterfactual supposes away is not a particular event or fact at all, but rather a *class* of them. Usually, counterfactuals that begin 'if *a* had not happened . . . ' really mean 'if neither a nor anything relevantly *a*-like had happened . . . ' I think that this is the right way to go, though I shall not defend it here.

[20] Lewis himself has recently said that when we imagine an event C away, "we imagine that C is completely and cleanly excised from history, leaving behind no fragment or approximation of itself" (2000, 190). I should emphasize, though, that my claim is just that there is a *presumption* against replacement; such evaluations are not always inappropriate any more than back-tracking ones are. Sometimes what we are interested in is precisely whether the closest worlds in which the antecedent is true are deletion worlds or replacement worlds—for example, when we want to know whether a backup system or conspiracy is in place. Normally, though, we ignore replacement worlds.

[21] In the current context, then, the presumption against replacement and the presumption against backtracking have rather different motivations. The ban on backtracking needs to be in place to make the counterfactuals properly function as a necessary condition on overdetermination. The ban on replacement emerges from the distinct requirement that c_1 and c_2 be causally sufficient for e.

[22] Kim has recently suggested that what amounts to what I have been calling the truth of (O1) on the deletion reading would undermine completeness (1998, 45). This is a bit of an overstatement. It certainly does not entail that completeness is false at the *actual* world, just in the world relevant to evaluating the counterfactual. And that is not obviously a problem—most people think physicalism is contingent, and if *it* is, so is completeness. So the issue has to instead be that the truth of (O1) on a deletion reading renders completeness *too* contingent, that it entails that completeness fails in inappropriately nearby possible worlds. But it turns out that whether or not this is the case depends upon how the presumption against replacement

evaluations is accommodated. Treating this fully is a task for another paper, but I do not think that the worry about the falsity of completeness has much bite.

[23] Using the insistence on deletion to argue for vacuity does look more promising for the event version of (O1). It seems more promising to argue that although it is not impossible for some particular event *m* to happen without *p*, it is indeed impossible for *that very event m* to occur without *any* relevantly *p*-like event. The claim would simply be that *m* is essentially physically realized, even though it is not essentially realized *by p*. So why not say that the event version of (O1) actually *is* vacuous, when interpreted properly—i.e., by deletion?

The answer is that this line misses the point behind the insistence on deletion evaluations. The insistence on evaluation-by-deletion rather than evaluation-by-replacement basically amounts to the claim that the truth value of (O1) should not be held hostage to facts about the essence of *p*, the event that is supposed away. That is, it is a bad idea to make the minimal change necessary to violate *p*'s essence, and, since that is a world in which *p* does not occur, conclude that we have reached a world where the antecedent of the counterfactual holds. But why should this quite general idea about how to evaluate counterfactuals only apply to the event that is supposed *away*, rather than also to the event explicitly held constant? Why think that the truth-value of the counterfactual is beholden to facts about the precise essence of *m* when it is not so beholden to facts about the precise essence of *p*? I do not think we should think that. If we *did*, we would have to conclude that a great many counterfactuals that are intuitively *false* are instead vacuously true. (Consider the following counterfactual about my microwave: 'if this humming noise had occurred without this pattern of electronic activity, my food would still have gotten hot'.) So (O1) is not properly understood as saying 'if *m* had occurred without anything relevantly *p*-like, *e* would still have occurred.' It is instead properly understood as saying 'if a relevantly *m*-like event had occurred without anything relevantly *p*-like, *e* would still have occurred'. And *that* is no more vacuous than is the straightforwardly property version of the counterfactual. Given the contingency of physicalism, a relevantly *m*-like event could occur without any physical event—even if *m* itself could not.

[24] I have purposely left constitution off the list, despite the fact that some—notably Pereboom and Kornblith 1991—want to deploy the claim that the physical constitutes the mental to solve the exclusion problem. Constitution cannot render (O2) vacuous, because it is not an upward necessitation relation. The fact that *a* constitutes *b* does not mean that *b* occurs in every metaphysically possible world in which *a* does—consider the standard case of a lump of clay that constitutes a statue. Constituted things can have lots of properties essentially that their constituting matter has only accidentally (for more, see Thomson 1998).

[25] The only other argument of this kind against token identity that I have actually found in the literature is Kripke's argument that a particular brain event that in fact underlies a pain could exist without any phenomenal character at all (1980, 146). But further such arguments can easily be made. For example, one can be run assuming some kind of conceptual role semantics, rather than the externalism in the main text. The difference is just that on such a view, a belief that ϕ winds up essentially bearing certain relationships to other of the subject's mental states, rather than to things out in the external world. A very similar argument can be run from functionalism about *non*contentful mental states, and I am quite sure that there are others.

[26] One upshot of this discussion is that physical events are not related to mental events as determinate to determinable, as Yablo would have it (1992).

[27] Depending on your view about the nature of events—a matter about which I have purposely remained silent—you might think that nothing more needs to be said here. If events are Kimian property exemplifications, it looks like the existence of necessitation relations between mental and physical properties would entail necessitation relations between instantiations of those properties—i.e., between mental and physical events. So if Kim is right about what events are, and if I am right that the relevant necessitation relations do *not* hold between the events, it might look like I have already established that they do not hold between the properties, either. This is not the case; making the connection requires assuming that events

are *essentially* instantiations of whatever property they are actually instantiations of. Kim himself denies this (1976, 47–48).

[28] This worry has more force in the token case, but I do think it applies to properties as well.

[29] It might be objected that this is an excessively extreme example. Perhaps we do not need to go all the way to a world in which *p* (P) occurs in a *vat* in order to get it to occur without *m* (M). What if *m* is a desire for water, and *e* the act of turning on a faucet? Given externalism, it looks as though the closest world in which *p* occurs without *m* is a world in which the watery stuff is XYZ—and in *that* kind of world, *p* would still cause me to turn on a faucet. Not so; do not forget the ban on replacement evaluations, which make counterfactuals with negative antecedents true too easily. The closest world relevant to the evaluation of (O2) is not one in which *p* occurs without *m* but does occur with something suspiciously *m*-like—here, a desire for twater—but one in which *p* occurs without anything relevantly *m*-like at all. And *that* kind of departure from actuality is quite unlikely to be one in which *p* causes the faucet to be turned on.

References

Balog, Katalin. 1999. "Conceivability, Possibility, and the Mind-Body Problem," *The Philosophical Review* 108: 497–528.

Bennett, Jonathan. 1987. "Event Causation: the Counterfactual Analysis," reprinted (1993) in Ernest Sosa and Michael Tooley, eds., *Causation*. Oxford: Oxford University Press, 217–233.

Blackburn, Simon. 1991. "Losing Your Mind: Physics, Identity, and Folk Burglar Protection," reprinted in *Essays in Quasi-Realism*. Oxford: Oxford University Press, 229–254.

Block, Ned. 1990. "Can the Mind Change the World?" in *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge: Cambridge University Press, 137–170.

Block, Ned, and Stalnaker, Robert. 1999. "Conceptual Analysis, Dualism, and the Explanatory Gap," *The Philosophical Review* 108: 1–46.

Boyd, Richard. 1980. "Materialism Without Reductionism: What Physicalism Does Not Entail," in Ned Block (ed.), *Readings in the Philosophy of Psychology*, *Vol. I*. Cambridge, MA: Harvard University Press.

Burge, Tyler. 1993. "Mind-Body Causation and Explanatory Practice," in J. Heil and A. Mele, eds., *Mental Causation*. Oxford: Clarendon Press, 97–120.

Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Davidson, Donald. 1970. "Mental Events," reprinted (1980) in *Essays on Actions and Events*. Oxford: Clarendon Press, 207–225.

Dretske, Fred, and Snyder, Aaron. 1973. "Causality and Sufficiency: Reply to Beauchamp," *Philosophy of Science* 40: 288–291.

Fodor, Jerry. 1989. "Making Mind Matter More," reprinted (1990) in *A Theory of Content and Other Essays*. Cambridge, MA: Bradford, 137–159.

Goldman, Alvin. 1969. "The Compatibility of Mechanism and Purpose," *The Philosophical Review* 78: 468–482.

Horgan, Terence. 1987. "Supervenient Qualia," *The Philosophical Review* 96: 491–520.

——— 1997. "Kim on Mental Causation and Causal Exclusion," *Philosophical Perspectives* 11: 165–184.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Clarendon.

Jackson, Frank, and Pettit, Philip. 1990a. "Causation in the Philosophy of Mind," *Philosophy and Phenomenological Research* 50: 195–214.

——— 1990b. "Program Explanation: A General Perspective," *Analysis* 50: 107–117.

Kim, Jaegwon. 1976. "Events as Property Exemplifications," reprinted (1993) in *Supervenience and Mind*. Cambridge: Cambridge University Press, 33–52.

——— 1989a. "Mechanism, Purpose, and Explanatory Exclusion," reprinted (1993) in *Supervenience and Mind*. Cambridge: Cambridge University Press, 237–264.
——— 1989b. "The Myth of Nonreductive Physicalism," reprinted (1993) in *Supervenience and Mind*. Cambridge: Cambridge University Press, 265–284.
——— 1993a. "The Nonreductivist's Troubles with Mental Causation," reprinted (1993) in *Supervenience and Mind*. Cambridge: Cambridge University Press, 336–357.
——— 1993b. "Postscripts on Mental Causation," in *Supervenience and Mind*. Cambridge: Cambridge University Press, 358–367.
——— 1995. "Mental Causation in Searle's 'Biological Naturalism'," *Philosophy and Phenomenological Research* 55: 189–194.
——— 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: Bradford.
Kripke, Saul. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Lepore, Ernest, and Loewer, Barry. 1987. "Mind Matters," *The Journal of Philosophy* 84: 630–642.
Lewis, David. 1973a. "Causation," reprinted (1986) in *Philosophical Papers*, *Volume II*. NY: Oxford, 159–213.
——— 1973b. *Counterfactuals*. Oxford: Basil Blackwell.
——— 1979. "Counterfactual Dependence and Time's Arrow," reprinted (1986) in *Philosophical Papers*, *Volume II*. Oxford: Oxford University Press, 32–66.
——— 1983. "New Work for a Theory of Universals," *Australasian Journal of Philosophy* 61: 343–377.
——— 2000. "Causation as Influence," *The Journal of Philosophy* 97: 182–197.
Loar, Brian. 1999. David Chalmers' *The Conscious Mind. Philosophy and Phenomenological Research* 59: 465–472.
Malcolm, Norman. 1968. "The Conceivability of Mechanism," *The Philosophical Review* 77: 45–72.
McDermott, Michael. 1995. "Redundant Causation," *The British Journal for the Philosophy of Science* 46: 523–544.
Mellor, D. H. 1995. *The Facts of Causation*. NY: Routledge.
Mills, Eugene. 1996. "Interactionism and Overdetermination," *American Philosophical Quarterly* 33: 105–117.
Noordhof, Paul. 1997. "Making the Change: the Functionalist's Way," *The British Journal for the Philosophy of Science* 48: 233–250.
Peacocke, Christopher. 1979. *Holistic Explanation*. Oxford: Clarendon Press.
Pereboom, Derk, and Kornblith, Hilary. 1991. "The Metaphysics of Irreducibility," *Philosophical Studies* 63: 125–145.
Pietroski, Paul M. 1994. "Mental Causation for Dualists," *Mind and Language* 9: 336–366.
Prior, Elisabeth, Pargetter, Robert, and Jackson, Frank. 1982. "Three Theses About Dispositions," *American Philosophical Quarterly* 19: 251–258.
Schiffer, Stephen. 1987. *Remnants of Meaning*. Cambridge, MA: Bradford.
Searle, John. 1992. *The Rediscovery of the Mind*. Cambridge, MA: Bradford.
Shoemaker, Sydney. 1975. "Functionalism and Qualia," *Philosophical Studies* 27: 292–315.
Thomson, Judith Jarvis. 1998. "The Statue and the Clay," *Noûs* 32: 149–173.
Yablo, Stephen. 1992a. "Cause and Essence," *Synthèse* 93: 403–449.
——— 1992b. "Mental Causation," *The Philosophical Review* 101: 245–280.
——— 1997. "Wide Causation," in James Tomberlin, ed., *Philosophical Perspectives 11*: 251–281.
——— 2000. "Textbook Kripkeanism and the Open Texture of Concepts," *Pacific Philosophical Quarterly* 81: 98–122.