Brit. J. Phil. Sci. **46** (1995), 523–544

Redundant Causation

Michael McDermott

ABSTRACT

I propose an amendment to Lewis's counterfactual analysis of causation, designed to overcome some difficulties concerning redundant causation.

This is a contribution to the counterfactual analysis of causation. Our starting-point is that, at least in the simplest kind of case, c is a cause of e (where c and e are distinct actual events) iff e would not have occurred if c had not occurred. Cases of redundant causation show that this will not do as a *general* analysis of causation. The question is how to generalize the simple analysis so that it will give the right answers for cases of redundant causation as well. The leading suggestion is David Lewis's.[1] I will criticize Lewis's account and propose an alternative.

Lewis thinks there are *other* cases where the simple analysis does not work, namely those involving chancy causation. I disagree,[2] but we will avoid that issue by making the simplifying assumption of determinism (as in Lewis's original article). There is also another kind of case which I see as counterexamples to the simple analysis but Lewis does not.[3] Such cases will likewise be excluded here.

Like Lewis, I accept that the simple analysis gives a sufficient condition for causation, but not a necessary condition.

Definition

Events c_1 and c_2 are *redundant causes* of e (c_1, c_2 and e being distinct actual events) under the following conditions: 'if either one of c_1 and c_2 had

[1] 'Causation' and Postscripts; Lewis [1986], Ch. 21.
[2] I think the simple analysis works just as well for chancy causation.
[3] McDermott [1995].

occurred without the other, then also e would have occurred; but if neither c_1 nor c_2 had occurred, then e would not have occurred.'

Lewis divides up redundant causation in the following way. If c_1 and c_2 have, intuitively, an equal claim to be regarded as causes of e, it is *over-determination*. If it is clear, intuitively, that one is a cause and the other merely waits in reserve, it is *preemption*.

1 Counterexamples to Lewis's analysis

Given determinism, Lewis says that e *causally depends* on c iff e would not have occurred if c had not occurred. Then, he says, c is a cause of e iff they are connected by a chain of causal dependence. (In the simple case, the chain has just one link.)

This account does give the right answers in *some* cases of redundant causation. Suppose I push Jones in front of a truck, which hits him and kills him; if I had not done so, he would have been hit and killed by a bus. Common sense says that my push was a cause of his death. But the death would have occurred without my push—the death was not causally dependent on the push. However, the push and the death are connected by a two-link chain of causal dependence: if the push had not occurred, Jones's collision with the truck would not have occurred, and if Jones's collision with the truck had not occurred, the death would not have occurred.

Objection 1: if the collision with the truck had not occurred, it would have been because the push had not occurred, and in that case the death *would* have still occurred, because Jones would have been hit by the bus.

Reply (this is Lewis's reply, and I endorse it): this is a 'backtracking' counterfactual, irrelevant to causation. What counts is what would have happened if I had pushed Jones out of the path of the bus and *then* the truck had somehow failed to hit him. We know it *would* have hit him in that case, but we can still ask what would have happened if it had not.[4]

Objection 2: the push did not cause Jones's death, because the death would have occurred whether or not the push occurred. The simple analysis gives the right answer, not the amendment.

Reply (Lewis does not consider this objection; I reply on his behalf): this is the initial intuition of quite a few competent speakers. But I have found that it can be easily corrected. I say, 'If the bus had not been there, and I

[4] Lewis has a complicated and controversial *theory* of counterfactuals, in terms of which he explicates the distinction between the 'standard' interpretation of counterfactuals and the one which permits the truth of backtrackers. We do not need to endorse all that. All we need accept is that a counterfactual like 'If the truck had not hit him, he would not have died' *has* two reasonable interpretations, one on which it is true, and one on which it is false. We can then stipulate that the reading for the counterfactuals used in the analysis of causal statements is to be the former.

had not pushed, Jones would not have died. So between us—me and the bus—we caused his death. *Which one* of us caused his death—me or the bus (or both together)?' And nearly everyone then retracts his initial intuition and says, 'Well, it must have been your push that did it—the bus clearly contributed nothing.'

(Perhaps you will doubt that the initial intuition I report here is common enough to merit attention. In cases of redundant *prevention*, however, which are perfectly parallel, the corrective is needed in the vast majority of cases. Suppose that I reach out and catch a passing cricket ball. The next thing along in the ball's direction of motion was a solid brick wall. Beyond that was a window. Did my action prevent the ball hitting the window? (Did it cause the ball to *not* hit the window?) Nearly everyone's initial intuition is, 'No, because it wouldn't have hit the window irrespective of whether you had acted or not.' To this I say, 'If the wall had not been there, and I had not acted, the ball would have hit the window. So between us—me and the wall—we prevented the ball hitting the window. *Which one* of us prevented the ball hitting the window—me or the wall (or both together)?' And nearly everyone then retracts his initial intuition and says, 'Well, it must have been your action that did it—the wall clearly contributed nothing.')

That was a case where Lewis's account gives the right answers. In general, however, it is not satisfactory. I will argue, first, that the existence of a chain of causal dependence between c and e is not *necessary* for the truth of 'c causes e'. The suggested counterexamples are of three kinds.

First, overdetermination without chains of causal dependence: I think there are cases where there is no chain of causal dependence from c_1 to e, or from c_2 to e, but common sense says that both c_1 and c_2 are causes of e. Lewis goes *some* way towards conceding this: he concedes that there are cases where common sense does not agree with the theory's verdict that c_1 and c_2 are definitely *not* causes. But he holds that in such cases common sense does not give a clear positive verdict either. It would be better if the analysis explained or reproduced the common-sense indecision as to whether c_1 and c_2 are causes of e or not, but we can live with an analysis which gives a definite 'no'.

Secondly, late preemption: Lewis picks out this kind of case by using the intuitive (unanalysed) idea of a series of events constituting a *causal process*. There is a completed causal process running from c_1 to e; there is no completed causal process running from c_2 to e; so c_1 is clearly a cause of e, c_2 clearly not. But the process from c_2 is not cut off until e occurs, so that e is not causally dependent on any event in the process from c_1—there is no chain of causal dependence from c_1 to e. (The events in the actual causal process from c_1 to e do not form a chain of causal dependence

because, for any intermediate event d in this process, e is not causally dependent on d—if d had not occurred e would not have occurred when it did, but it would have occurred later.)

Lewis says[5] there are plenty of cases of this kind, and they are counter-examples to his original analysis. He describes a possible amendment, but it is not clear whether he actually adopts it.[6] In any case, the amendment is also unsatisfactory, I will argue.

Thirdly, there are cases where c_1 is clearly a cause of e, and c_2 not, although in the absence of c_1 e would have occurred when it actually did occur, and there is no chain of causal dependence from c_1 to e. Lewis considers one such case.[7] He says that, although it would be better if the analysis did not disagree with the clear verdict of common sense, we can properly accept a theory which does, because cases of this kind are not physically realistic. I will argue that this reply is not satisfactory.

I will also argue that the existence of a chain of causal dependence is not *sufficient* for the truth of 'c causes e'. Since I am allowing that each link in the chain is a case of causation, this means that I will be arguing that causation is not transitive. My fourth group of cases, then, are supposed to be counterexamples to causal transitivity. Lewis does not consider the possibility of such cases—he simply takes transitivity for granted.

1.1 Overdetermination

Can we suppose that an analysis is fundamentally correct if it gives a definite negative verdict about cases where common sense is indecisive? That is a nice question. But I think the true problem for Lewis is much worse: there are actually cases where common sense gives a clear positive and the analysis gives a clear negative. I do not have in mind cases of a sort not considered by Lewis. My disagreement is that on some of the cases where Lewis says common sense is indecisive I reckon common sense gives a clear positive.

What we want are the intuitions of competent speakers uncorrupted by philosophical theory. Let me report a little practical experiment. I asked a group of naïve subjects (randomly selected first year undergraduates, who had been taught nothing remotely connected with the philosophy of causation) the following problem:

[5] Postscript E.

[6] Lewis says in Postscript E that the amended analysis 'may well be preferable' to the original. But when he summarizes his views on causation in a later chapter, it is the *original* analysis which he gives (p. 242).

[7] p. 202. (Rather confusingly, he calls it a case of late preemption.)

> Prof Jones is sitting in a metal chair. Two sets of wires attached to the chair lead to two power points. Mr A switches on the power at one point. At precisely the same moment Mr B (acting independently) switches on the power at the other point. Jones gets a fatal shock. He would have got exactly the same shock if either A or B had acted alone. Are the following statements true or false?
>
> Mr A caused Jones's death.
>
> Mr B caused Jones's death.

Of my 40 respondents (answering independently), 33 said that both statements were true.

Lewis might reply, however, that his theory can accommodate this positive verdict. In many cases of overdetermination, he holds, close inspection of the physical details reveals a 'Bunzl event', an event which is jointly caused, without redundancy, by c_1 and c_2, and which is itself a cause of e. In such cases c_1 counts as a cause of e because the sequence $< c_1, b, e >$ is a chain of causal dependence (and similarly for c_2). Perhaps Lewis could reply that the true microphysics of the case of Prof Jones discloses a Bunzl event, and hence enables his analysis to agree with common sense. Bunzl himself[8] suggests that if we consider the identities of the individual electrons whose passage through Jones's body constitutes the fatal shock, we will find that *that* event would not have occurred unless *both* A and B had acted as they did: if either had failed to act, a somewhat different group of electrons would have passed through Jones —he would not have really got 'exactly the same shock'.

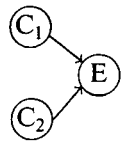
We must remember, however, that our project—and Lewis's—is conceptual analysis. The test of a theory in conceptual analysis is whether it yields the common-sense verdict, not on the *true* facts of the case, but on the facts of the case as our commonsensical informants take them to be. So Lewis can only make his theory agree with the common-sense verdict on the Jones case if he can show that the speakers who give this verdict *believe* there is a Bunzl event. The relevant question is: when asked to say *why* they think Mr A's act was a cause of Jones's death, do people *reject* the stipulation that he 'would have got exactly the same shock if either A or B had acted alone'? Do they talk about the identities of the individual electrons whose passage constituted the fatal shock? Do they refer to anything else which could conceivably play the role of a Bunzl event? And the answer is: no, they do not.

What they do say (when they say anything non-vacuous) is often along the following lines: A's act is a cause of Jones's death because there is a continuous causal process linking the two; similarly for B's act; both

[8] Bunzl [1979].

processes run to completion, neither cuts the other short; neither cause preempts the other, so both are genuine causes. There is no sign here of the idea that what makes the case genuine causation is a Bunzl event.

Here is a case of *direct* overdetermination, in which there is no possibility of a hidden Bunzl event:



Neuron E is doubly stimulated by the firing of C_1 and C_2, but fires exactly as if it had been stimulated by C_1 or C_2 alone.[9] Here Lewis's theory has no way to avoid a negative verdict: the firing of C_1, for example, was definitely not a cause of the firing of E. But my respondents favoured a positive verdict just as strongly as in the electrocution case.

1.2 Late preemption

Let us have a case to illustrate the problem. Two shots are fired in quick succession. The first bullet perforates Jones, the second passes through the hole in its wake, scarcely touching the sides. Clearly it is the first shot which causes his death. But there is no chain of causal dependence linking them. There is indeed a series of events which, intuitively, constitute a causal process: < trigger squeezed, first bullet leaves gun, first bullet arrives at Jones, first bullet passes through Jones, Jones dies > . But the final effect is not causally dependent on even the immediately preceding event in this series—because of the second shot, Jones would still have died (a moment later) if the first bullet had not passed through him.

Lewis explores two ideas which might solve the problem. The one he seems to favour is based on the following intuition: whether a series of events in a given spatiotemporal region is a causal process or not is an intrinsic matter—it depends just on what happens in that region (and on the laws of nature), not on what happens in its surroundings. Consider the series S of events linking the first shot with the death. It is not a chain of causal dependence, but it would be, if it were not for the second shot. Lewis thinks it likely that if we consider *all* the regions (of actual and possible worlds which share our laws) which have the same intrinsic character as the region in which S occurs, we will find that the majority do *not* have second

[9] Lewis seems undecided about how realistic the assumption is that there is a definite threshold for E's firing (namely, that *at least one* of C_1 and C_2 should fire)—compare p. 196 with p. 212.

bullets in their surroundings, or anything else which prevents the series being a chain of causal dependence; indeed in the majority the death will be causally dependent on the shot. If so, we say that the death *quasi-depends* on the first shot, in *all* the regions which share that intrinsic character. The revised analysis, then, says that c is a cause of e iff they are linked by a chain of *quasi*-dependence.

We might well wonder how we can be sure that in the majority of possible cases there is no second bullet. But that is a comparatively small problem. The big problem is that the basic idea behind the revised analysis—that whether a series of events in a given spatiotemporal region is a causal process or not is an intrinsic matter—is quite contrary to intuition. Here is a counterexample.

World w_1 is pretty much like the actual world. One day, in a fit of bad temper, Nixon reaches for the button. Instantly Haig leaps forward, throws Nixon to the ground, and sits on his head until he calms down. Next morning, Joe Blow eats breakfast as usual. If Haig had not leapt into action, Nixon would have pressed the button, triggering a nuclear attack on Russia; the Russian response would have destroyed all human life in America; Blow's breakfast would have remained uneaten. Haig's leap, then, was a cause of Blow's eating breakfast—the latter would not have occurred without the former. (The leap caused the breakfast by preventing something which would have prevented it.)

World w_2 shares the laws of w_1, and the relevant events in America are the same: Haig's leap prevents Nixon's pressing the button, and thereby prevents the launching of the nuclear attack; Blow eats breakfast next morning. But in w_2 Russia is entirely uninhabited. If the American attack had been launched, the bombs would have fallen harmlessly in the desert; there would have been no counterattack; Blow would have eaten his breakfast next morning undisturbed. It seems clear that in w_2 Haig's leap was *not* a cause of Blow's eating breakfast. The series of events <Haig's leaping, Blow's eating> is a causal process in w_1, it occurs in an intrinsically identical region in w_2, but it is not a causal process in w_2.

Lewis's other idea for solving the problem of late preemption is to postulate that the effect has an extremely rich essence. In particular its precise *time* is of the essence: if Jones, for example, had died at a slightly different time, that would not have been the same event as the death which actually occurred. On this assumption, the first shot caused Jones's death, because the death—the *actual* death—would not have occurred if the first shot had not occurred. Similarly for all cases of late preemption, since in all such cases the effect would have been delayed (as we say) if the preempting cause had not occurred.

The trouble is that this strategy also yields causal relationships where we

don't want them. Common sense tells us that many things affect the time at which an event occurs, but are not causes of it. For example, suppose that Jones's execution was scheduled for 6 a.m., but that a slight malfunction in the official alarm clock resulted in his being executed a minute or so later (or earlier). We would not say that the malfunction was a cause of his death. But that is how it turns out, if we give the death the rich essence needed to solve the problem of late preemption: the actual death would not have occurred in the absence of the malfunction.

You can get these spurious causal relationships even in cases of late preemption. Suppose that the second shot makes a particularly loud noise; this causes Jones to twitch slightly (while the first bullet is still on its way), so that the first bullet (with the second in its wake) perforates a somewhat different part of his body; this makes him die substantially more (or less) quickly. Without the first shot, he would have died at a *slightly* different time; without the second, he would have died at a *substantially* different time; if the first shot was a cause of his death because of the death's rich essence, so was the second. But common sense is still clear that the first shot was the sole cause of Jones's death.

For this kind of reason, Lewis rejects the 'rich essence' strategy (correctly, in my opinion). He evidently continues to feel, however, that it might turn out to have been on the right track[10] (and I think that is right, too; see below).

1.3 Preempting direct cause

In early preemption, the effect is not causally dependent on the preempting cause. Lewis makes it turn out to be a genuine cause by appealing to intermediate events which will link cause and effect in a *chain* of causal dependence. But, many philosophers have asked, can't a preempting cause produce its effect *directly*?

Suppose that at 11 p.m. Witch A performs a spell to make the Prince turn into a frog at midnight. We stipulate that the spell works directly, without intermediate events. If Witch A had not done so, Witch B would have performed at 11:30 p.m. a spell to make the Prince turn into a frog at midnight. The Prince turns into a frog at midnight, and would have done so in the absence of spell A, though not in the absence of spell B as well. Common sense says that this is a case of early preemption: spell A actually caused the transformation, although it would still have occurred, with a different cause, in the absence of spell A.

If we accept the common-sense verdict on this case, it is one where there is a temporal gap between cause and effect, not filled by any causal process.

[10] p. 205.

Lewis discusses a case with such a temporal gap, although the subject matter is Lewis's usual one, the firings of neurons. He grants that the case is logically possible, and that common sense gives a clear positive verdict, despite the absence of a chain of causal dependence. But he says that agreement with common sense in such cases is 'not an urgent goal' because they 'go against what we take to be the ways of this world; they violate the presuppositions of our habits of thought ...'.[11] What Lewis means, I take it, is that our common-sense judgements are very likely to be influenced by our ingrained knowledge that temporal gaps with nothing going on *never* occur in the workings of neurons, despite the stipulation of the case. It was precisely to avoid this response that I took my example from the realm of fairy tale. The example is instantly comprehensible, with the 'common-sense' judgements of causation, by young children. It would be quite implausible to argue that their judgements are unconsciously based on an ingrained belief that the processes by which witches' spells are implemented never involve temporal gaps.

1.4 Counterexamples to transitivity

(i) *The dog-bite.* My dog bites off my right forefinger. Next day I have occasion to detonate a bomb. I do it the only way I can, by pressing the button with my left forefinger; if the dog-bite had not occurred, I would have pressed the button with my right forefinger. The bomb duly explodes. It seems clear that my pressing the button with my left forefinger was caused by the dog-bite, and that it caused the explosion; yet the dog-bite was not a cause of the explosion.

Can Lewis meet the objection by appealing to a 'profligate' theory of events? Suppose we say (as Lewis sometimes suggests)[12] that *two* actual button-pressings occurred, with different essences: strong was essentially a pressing with the left forefinger, and would not have occurred if I had used the right; weak was only accidentally a pressing with the left forefinger, and would have still occurred if I had used the right. Can we then say that neither of these simultaneous button-pressings was *both* caused by the dog-bite and a cause of the explosion? No; for Strong *was* both caused by the dog-bite and a cause of the explosion. Strong was a cause of the explosion because if it had not occurred—if I had not pressed the button with my left forefinger—I would not have, because I could not have, pressed it at all. (*Objection*: if you hadn't pressed it with the left you *would* have pressed it with the right, because in that case here would have been no dog-bite. *Reply*: this is a 'backtracking' counterfactual, irrelevant to causation.)

[11] p. 203.
[12] 'Events', §IV.

The crucial feature of the example is that the dog bit the finger right off. If the finger had only been injured, it would not have been clearly true that if I had not pressed the button with the left forefinger I would not have pressed it at all. Then we would have merely had a counterexample to transitivity-plus-coarse-grained-theory-of-events.[13] (Although that would be bad enough for transitivity: even Lewis seems to find the profligate theory of events an embarrassment.)[14]

(ii) '*Shock C*' is a game for three players. A and B each has a switch with two positions, Left and Right; to start, both are in the Left position. A has first turn: he can either move his switch to Right, or do nothing. B then has a turn: he can either move his switch to Right, or do nothing. The power is then turned on: if both switches are in the Left position, or both in the Right position, C gets an electric shock.

On this occasion the play goes as follows. A moves his switch to Right. B observes A's move; he wants C to get a shock, so he responds by moving his switch to Right also. C duly gets a shock.

Common sense tells us that A's move was a cause of B's move, that B's move was a cause of C's shock, but that A's move was *not* a cause of C's shock.

Among naïve subjects, this seems to be the most robust of my three counterexamples to transitivity. 80% of my sample gave the stated verdicts. (Of the rest, some denied that A's move was a cause of B's move, some denied that B's move was a cause of C's shock, and some said that A's move *was* a cause of C's shock.)

(iii) *The massage*. I give Jones a chest massage, saving his life: without the massage he would have died in minutes. When he recovers, Jones goes to New York, where he eventually meets a violent death. It seems that the massage was a cause of his going to New York, his going to New York was a cause of his death, but the massage was not a cause of his death.

The intuition that the massage was not a cause of his death can be suppressed if the case is confused with one of ordinary preemption, like the following: I advise Jones to go to New York; as a result he goes, and years later meets a violent death there; shortly after his departure a tree falls on his old home, which would certainly have killed him if he had not gone to New York. Here it seems that my advice caused his trip, his trip caused his death, and my advice *did* cause his death.

In both cases Jones would have died whether or not I acted as I did. In both cases my action *delayed* his death. The difference is that in the

[13] Like those of Hausman [1992] and Ehring [1987].

[14] Its falsity seems to be taken for granted throughout Postscript E to 'Causation': Lewis's line there is that whether the actual button-pressing would have occurred in the absence of the dog-bite depends on 'standards of fragility' which are 'both vague and shifty' (p. 197); whereas on the profligate theory there is no such thing as *the* actual button-pressing.

massage case my saving him from the early death was a *means* to his dying later; the massage caused him to go to New York *by* causing him not to die early. If he had not been in any real danger, so that my massage did not really save his life, the massage would *not* have been a cause of his going to New York. Whereas in the other case my saving Jones from the early death was merely a *by-product* of the chain of events which actually led to his death. If there had been no falling tree, so that my advice did not really save his life, the advice *would* still have been a cause of his going to New York.

When this difference between the two cases is made clear, it is natural to say that my action did not cause his death in The Massage, although it did in The Advice. Common sense tells us that you can't cause Jones's death by a process whose efficacy depends on its prolonging his life, though you can cause his death by an action which *incidentally* closes off an alternative route to his death.

2 The alternative analysis

In presenting my alternative counterfactual analysis of causation, I will proceed in three stages. To begin with, we concentrate on events which would not have occurred at all if they had not occurred when and how they actually did. In the first stage, we deal with *direct* causation. In the second, we ask when a causal *chain* amounts to a relation of cause and effect. Finally we turn to things which affect the time or manner of occurrence of an event, and see how these results may be used to resolve cases of late preemption.

2.1 Direct causation

Let us explore the intuitive idea that a cause is a *part of a minimal sufficient condition*. What does it mean, then, to say that the occurrence of a set of events C is a *sufficient condition* for the occurrence of e? Usually this is analysed in terms of natural laws, but what we want is an analysis in terms of counterfactuals. I suggest this: given C, e would have occurred even if any other actual events had failed to occur; that is, no matter what other actual events had failed to occur, e would still have occurred as long as the members of C did.

(D1) C is a sufficient condition for e

iff $\sim(\exists D)(e \notin D \wedge (\forall x)(x \in C \supset O(x)) \wedge (\forall x)(x \in D \supset \sim O(x))$ $\Box\!\!\rightarrow \sim O(e))$

[i.e. iff there is no set of actual events D, apart from e itself, such that (if the members of C had all occurred and the members of D had all failed to occur, then e would have failed to occur)].

Provisionally defining a *minimal* sufficient condition as one which has no sufficient condition as a proper subset, let us see how this works on a

couple of examples. We had a case of an (early) preempting direct cause—the spell of Witch A turning the Prince into a frog. The Prince's transformation was not causally dependent on the spell, because of the backup witch waiting in reserve. But spell A by itself was a sufficient condition for the transformation, and clearly a minimal sufficient condition. If a cause is a part of a minimal sufficient condition, spell A caused the transformation, as desired. We had a case of overdetermining direct causes—the neuron E doubly stimulated by C_1 and C_2. The firing of E was not causally dependent on the firing of C_1 (for example). But the firing of C_1 by itself was a sufficient condition for the firing of E, and clearly a minimal sufficient condition. If a cause is a part of a minimal sufficient condition, the firing of C_1 caused the firing of E, as desired.

(Or perhaps we should take a wider view of these cases. Perhaps we have been taking for granted certain cooperating causal factors, such as the Prince's still being alive at midnight: without that, he presumably would not have turned into a frog. If so, the minimal sufficient condition for the transformation should contain these events as well as spell A. But on the wider view spell A was still part of a minimal sufficient condition for the transformation. So let us keep things as simple as we can.)

So the notion of minimal sufficient condition seems a more promising key to direct causation than the notion of causal dependence. But we need to attend more closely to the definition of *minimality*. Consider the following case.

I say to the Prince 'What is my name?' Witch A has put a spell on him which will make him turn into a frog if he says 'Rumpelstiltskin'. Witch B has put a spell on him which will make him turn into a frog if he does *not* say 'Rumpelstiltskin'. He says 'Rumpelstiltskin' and turns into a frog. (As before, we stipulate that there are no intermediate events to provide chains of causal dependence.)

The common-sense causal judgements about this case are clear: spell A ('A') and the Prince's saying 'Rumpelstiltskin' ('R') were joint causes of his turning into a frog ('T'); spell B ('B') was not a cause. It is clear that {A,R} was a sufficient condition for T, A alone was not, R alone was not; and we are happy to say that {A,R} was a minimal sufficient condition. But what about {A,B}? It was a sufficient condition for T; we do not need to include R, since the Prince would have turned into a frog even if he had not said 'Rumpelstiltskin', as long as spell A and spell B had both occurred. And neither A nor B alone was a sufficient condition. But {A,B} was not a complete cause of T, so it would seem intuitively wrong that it should turn out to be a minimal sufficient condition.

The remedy, however, is also clear enough on intuitive grounds: if we consider the two possibilities R and not-R, we see that {A,B} contains a

redundant element *either way*: {A,B,R} would still be sufficient for T without B, and {A,B, not-R} would still be sufficient for T without A. To support the intuition that a complete cause is a minimal sufficient condition, then, we need a definition along the following lines:

(D2) *C* is a minimal sufficient condition for *e*
iff (i) *C* is a sufficient condition for *e*;
(ii) no proper subset of *C* is a sufficient condition for *e*;
(iii) there is no actual event *r* (distinct from *e* and the members of *C*) such that
—for some proper subset *C** of *C*, *C**∪{*r*} is a sufficient condition for *e*; and
—for some proper subset *C** of *C*, *C**∪{*not-r*} would be a sufficient condition for *e*.

This gives the desired results: {A,R} is a minimal sufficient condition for T, {A,B} is not. And it does not upset our provisional results in the earlier cases: the preempting spell A is still a minimal sufficient condition for the Prince's turning into a frog, and the firing of C_1 is still a minimal sufficient condition for the firing of E.

(Let us check that condition (iii) does not rule out {A,R} as a minimal sufficient condition for T. The only relevant candidate for the event *r* is spell B. Then we have that for the subset {A} of {A,R}, {A,B} is a sufficient condition for T; but neither {A, not-B} nor {R, not-B} would be a sufficient condition for T.)

I have not been able to find any intuitive counterexamples to the thesis that a part of a minimal sufficient condition is always a cause. The converse, however, is obviously false. Consider the simple case where, intuitively, *c* is a cause of *e*, and the relation depends on an intermediate event *d*: *c* actually caused *d*, which in turn caused *e*; and if *d* had somehow failed to occur, *e* would have failed to occur, despite the occurrence of *c*. Then *c* is a cause of *e*, but not a minimal sufficient condition for *e*, because it is not a sufficient condition for *e*. Nor is *c part* of a minimal sufficient condition for *e*; for the additions necessary to make it sufficient—*d* and its cooperating causal factors—would be sufficient by themselves. The hypothesis I will adopt is therefore only that *c* is a *direct* cause of *e* iff it is part of a minimal sufficient condition for *e*.

(D3) *c* is a direct cause of *e*
iff *c* is a member of some minimal sufficient condition for *e*.

2.2 Causal processes

Given sufficient information about the truth values of counterfactuals linking the relevant events, we can now determine which ones *directly*

cause which. The next step is to show how the facts of direct causation determine which events *cause* which.

Let us say that

(D4) there is a causal process from c to e
iff there are events $c', c'', \ldots, d_1, d_1', \ldots, d_n, d_n', \ldots$ such that
$\{c, c', \ldots\}$ is a minimal sufficient condition for d_1,
$\{d_1, d_1', \ldots\}$ is a minimal sufficient condition for d_2,
...
$\{d_n, d_n', \ldots\}$ is a minimal sufficient condition for e.

For later reference, let P be the set of all the events involved in the process; let Pi be the set of all the *intermediate* events $d_1, d_2, \ldots$; and let $P_c = P \cap (\overline{Pi \cup \{e\}})$ be the set of all *primary causal factors* involved in the process (the events which enter into the process as causes and not as effects).

Now it seems clear that c is a cause of e only if there is a causal process (of one or more steps) from c to e. But our counterexamples to the transitivity of causation seem to show that the converse is not true. Take the case of the game Shock C. We have a causal process as follows:

$$\left.\begin{array}{l}\text{A sw Right}\\ \text{B's desire}\end{array}\right\} \longrightarrow \left.\begin{array}{l}\text{B sw Right}\\ \text{A sw Right}\\ \text{general set-up}\end{array}\right\} \longrightarrow \text{shock.}$$

A's switching Right and B's desire for C to get a shock are joint direct causes of B's switching Right;[15] A's and B's switching Right, in conjunction with the general set-up of the electrical connections and so on, are joint direct causes of C's getting a shock. But A's switching Right is not a cause of C's getting a shock.

Why not? One relevant fact is clear: the shock would have occurred without A's act. But that cannot be all there is to it. In all cases of redundant causation c causes e although e would have occurred without c. We note also that in Shock C, if A's act had not occurred, there would have been a causal process *from A's not acting* to the shock:

$$\left.\begin{array}{l}\text{A does nothing}\\ \text{B's desire}\end{array}\right\} \longrightarrow \left.\begin{array}{l}\text{B does nothing}\\ \text{A does nothing}\\ \text{general set-up}\end{array}\right\} \longrightarrow \text{shock.}$$

This also fails to settle the matter. In preemption (though not in overdetermination) there would likewise be a causal process from *not-c* to e, if c failed to occur. Recall the case where I push Jones in front of a truck, which

[15] As before, I am artificially simplifying. On a slightly more realistic view, A's switching Right, in conjunction with the light being on, etc. causes events in B which, in conjunction with his desires and his beliefs, cause his own switching Right. But recognizing extra steps in the causal process would not change things in principle.

hits him and kills him; if I had not done so, he would have been hit and killed by a bus. We have the actual causal process:

$$\left.\begin{matrix}\text{push}\\ \text{truck}\end{matrix}\right\}\longrightarrow \text{hit by truck}\left.\begin{matrix}\\ \text{fragility}\end{matrix}\right\}\longrightarrow \text{death.}$$

My push and the presence of the truck are joint causes of his being hit by the truck; this and the fragility of his body and so on jointly cause his death. If I had not pushed we would have had:

$$\left.\begin{matrix}\text{no push}\\ \text{bus}\end{matrix}\right\}\longrightarrow \text{hit by bus}\left.\begin{matrix}\\ \text{fragility}\end{matrix}\right\}\longrightarrow \text{shock.}$$

My failure to push and the presence of the bus would have jointly caused his being hit by the bus; this and his fragility would have caused his death.

The crucial point, I suggest, can be put intuitively as follows: if there is just one actual causal process from c to e, but *not-c* would have led to e *by that process*, then c is not a cause of e. It seems natural to say that A's doing nothing would have led to the shock by the same process as A's switching Right did, whereas my failure to push would have led to the death by a different process than my push did. And why does this seem natural? *Not* because the alternative process would have gone by different *intermediate* events—that is true in both cases. The reason, I suggest, is that the alternative process would have depended on no *primary causal factors* beyond those of the actual process. If I had not pushed, the other events which actually cooperated in producing his death would not have been sufficient to produce his death; the presence of the bus, which played no part in the actual causal process, would have become essential. Whereas the factors which actually cooperated with A's switching Right to produce C's shock would still have been sufficient if A had not switched Right.

Let me now try to make this more precise. I will say that a set of events C is sufficient for e *relative to a process* P iff, given C, e would have occurred even if any other primary causal factors in P, or any events outside P, had failed to occur; that is, no matter what other primary causal factors in P, or any events outside P, had failed to occur, e would still have occurred as long as the members of C did.

(D5) C is a sufficient condition for e relative to P

$$\text{iff } \sim(\exists D)(D \subset ((P_c \cap \bar{C}) \cup \bar{P}) \wedge ((\forall x)(x \in C \supset \mathrm{O}(x)) \wedge (\forall x)(x \in D \supset \sim \mathrm{O}(x)) \Box\!\!\rightarrow \sim \mathrm{O}(e)))$$

[i.e. iff there is no set of actual events D, selected from the other primary causal factors in P, and events outside P, such that (if the members of C had all occurred and the members of D had all failed to occur, then e would have failed to occur)].

The notion of *minimal* sufficient condition may then be relativized to P in the obvious way:

(D6) C is a minimal sufficient condition for e relative to P
iff (i) C is a subset of P_c;
(ii) C is a sufficient condition for e relative to P;
(iii) no proper subset of C is a sufficient condition for e relative to P;
(iv) there is no actual event r (distinct from e and the members of C) such that
—for some proper subset C^* of C, $C^* \cup \{r\}$ is a sufficient condition for e relative to P; and
—for some proper subset C^* of C, $C^* \cup \{not\text{-}r\}$ would be a sufficient condition for e relative to P.

My (provisional) hypothesis, now, is that c causes e iff there is a causal process P from c to e and c is part of a minimal sufficient condition for e relative to P.

Let's apply this to the cases. In The Push, P is {my push, the presence of the truck, its hitting Jones, his fragility, his death}; P_c is {push, truck, fragility}. Did my push cause Jones's death? That is, is there a set C which is a minimal sufficient condition for the death relative to P, and which has the push as a member? Yes: C is exactly P_c. P_c is a sufficient condition for the death, relative to P, because as long as the push, the presence of the truck, and the fragility of Jones all occurred, Jones would die—whatever actual events outside P might have failed to occur. (If the truck had somehow failed to hit Jones, despite my push and the presence of the truck, he would not have died; but his being hit by the tuck is not outside P.) Furthermore, no proper subset of P_c is sufficient for Jones's death, relative to P. {truck, fragility} is not, in particular; for, letting D be {push, bus} (this is legitimate—the push is not a member of {truck, fragility}, and the bus's presence is not a member of P), we note that if the truck had still been present, and Jones had still been fragile, but there had been no push and no bus, Jones would not have died. Nor is there any other reason to think that P_c is not a *minimal* sufficient condition for the death, relative to P. So the push was a cause of the death.

In the Shock C case, P is {A sw Right, B's desire, B sw Right, general set-up, shock}; P_c is {A sw Right, B's desire, general set-up}. Did A's switching Right cause C's shock? That is, is there a set C which is a minimal sufficient condition for the shock relative to P, and which has A's switching Right as a member? No. The only plausible candidate for C is P_c, and in this case P_c is *not* a minimal sufficient condition for the shock, relative to P. For its proper subset {B's desire, general set-up} is a sufficient condition for the shock, relative to P. Given B's desire for C to get a shock, and the general

set-up, then even if A had done nothing, C's shock would have still occurred (irrespective of any events outside *P*).

Similarly for The Dog-bite. The bite, my desire for the bomb to explode, and the button–bomb connection were the primary causal factors in a process leading (via my pressing the button with my left forefinger) to the explosion. {Bite, desire, connection} was a sufficient condition for the explosion, relative to this process, but not a minimal sufficient condition; for the desire and the connection would have led to the explosion even if there had been no bite (though via a different intermediate event, a pressing with the right forefinger).

So the analysis agrees with the intuitive verdicts on these cases. It also allows for *direct* preemption, as we saw. (The definitions obviously imply that a direct cause is a cause.) Furthermore it gives the intuitive verdict in cases of overdetermination—namely, that the overdetermining causes are genuine causes, whether or not there is a Bunzl event.

In all these cases the current analysis gives the right results. It needs modification to cope with *late* preemption, but the final analysis will be equivalent to the current one except where the causal factors in question affect when or how the prospective 'effect' occurs, and not just whether it occurs or not.

2.3 Causing and affecting

Bernard made a trip across the Channel. The cause of his trip was that he had business in Boulogne. Bernard missed the ferry, so he swam. His missing the ferry did not *cause* his trip, but it *affected* his trip—it caused it to occur in one way rather than another. What is the basis for this distinction between causing and affecting?

The key is that the distinction is not 'in the objects', but in the way we refer to them. We find it natural to say that Bernard's *swim* was a result of his missing the ferry, but his *trip* was not. His swim would not have occurred if he had caught the ferry, but his trip still would have. And yet his trip *was* his swim: he only made one trip across the Channel, and it was a swim.

We can sort the matter out with the help of a little technical vocabulary. Let us say that every event has both a *real* essence and a *nominal* essence. Two possible events can have the same nominal essence, but not the same real essence. The nominal essence depends on how the event is referred to; the real essence does not. The nominal essence can be read off from the intrinsic features of the event used to refer to it. If we are talking about Bernard's trip, i.e. his swim, *as* a swim, its nominal essence is given by the sentence 'Bernard swam'; if we are talking about it as a trip its nominal

essence is given by 'Bernard took a trip'. The real essence is discovered by empirical investigation. It may turn out, for example, that the real essence of his trip, i.e. his swim, is given by the sentence 'Bernard swam across the Channel from Dover to Boulogne; he set off at 7:15a.m. in high spirits, a striped swimsuit and a south-easterly direction; for the first hour he swam overarm; then ...'. The real essence is, roughly speaking, a full intrinsic description of the event, including a precise specification of its time of occurrence. Time is of the *real* essence.

If we temporarily leave redundant causation to one side, we can now distinguish causing from affecting as follows. If e would not have occurred in the absence of c—if its *real* essence would not have been realized—then c either caused or affected e. If in addition no event with the same nominal essence as e would have occurred in the absence of c —if its nominal essence would not have been realized—then c counts as a cause of e. That is to say, c is a cause of e *under that description*. 'Missing the ferry caused Bernard's swim' is true because he would not have swum if he had caught the ferry. 'Missing the ferry caused Bernard's trip' is false because he would still have made a trip if he had caught the ferry.

We had an example where something affected an event's *time of* occurrence—the clock malfunction which delayed Jones's execution by one minute. The nominal essence of the execution is given by 'Jones was executed', the real essence by 'Jones was executed at 6:01 ... '. Without the clock malfunction there would have been an event with the same nominal essence, but not the same real essence. So 'The malfunction *caused* Jones's execution' is false, and 'The malfunction *affected* Jones's execution' is true.

Another example. Suppose that Frank's tension causes him to serve badly—without the tension he would still have served, but not badly. Was his serve, i.e. his *bad* serve, caused by tension? Without the tension, the real essence of the event in question would not have been realized; neither would there have been an event with the nominal essence determined by 'Frank's bad serve', though there would have been an event with the nominal essence determined by 'Frank's serve'. So 'Tension caused Frank's serve' is false, but 'Tension caused Frank's bad serve' is true; so is 'Tension affected Frank's serve'.

I said that nominal and real essences are intrinsic, not relational. In particular, an event's causes and effects are not parts of its nominal essence. Consider 'The passenger's failure to wear a seat belt caused the fatal accident'. The nominal essence of the alleged effect is 'There was an accident', not 'There was a fatal accident'. So the causal sentence is presumably false, although the passenger's failure was the cause of the accident's being a fatal accident.

The concept of 'real essence' will serve our purposes whether or not the *term* is appropriate. But actually the idea that events really have very rich essences has some good independent support, at least with regard to time of occurrence. Against it is the fact that we say '*this* event would have occurred at a different time if ....'. But we can easily convince ourselves that such sayings are not true, 'strictly speaking'. In world w_1 Jones is executed at 6:00; in w_2 (our world) he is executed at 6:01; is that the same event, delayed, or a different event? Well, consider world w_3, in which Jones is executed at 6:00, quickly resuscitated (by medical or divine intervention, with any wounds being perfectly repaired), and executed again at 6:01. Certainly there are *two* executions in w_3, distinguished solely by their times of occurrence. It seems obvious, also, that only the first occurs in w_1, and only the second in w_2. But that implies that the w_1 execution is a different event to the w_2 execution.

Returning now to the general case, the required amendment to our provisional analysis would seem to be as follows: c is a cause of e iff there is a causal process P from c to e—i.e. to the realization of the *real* essence of e—and c is part of a minimal sufficient condition for the realization of the *nominal* essence of e, relative to P. Where 'NO(e)' says that there occurs an event with the same nominal essence as e, we may define

(D7) C is a sufficient condition for the realisation of the nominal essence of e relative to P

iff $\sim(\exists D)(D \subset ((P_c \cap \bar{C}) \cup \bar{P})$
$\wedge ((\forall x)(x \in C \supset \mathrm{O}(x)) \wedge (\forall x)(x \in D \supset \sim \mathrm{O}(x)) \Box\!\!\rightarrow \sim \mathrm{NO}(e)))$.

The obvious amendment to (D5) gives us our definition of 'C is a minimal sufficient condition for the realisation of the nominal essence of e relative to P'. And then, finally,

(D8) c is a cause of e

iff there is a causal process P from c to e , and c is a member of a minimal sufficient condition for the realisation of the nominal essence of e relative to P.

Let us apply this analysis to our first case of late preemption, where Jones was killed by the first of two shots fired in quick succession. We want 'The first shot caused his death' to come out true, 'The second shot caused his death' false. There was a causal process leading from the *first* shot to the death, a process in which the second shot played no part. The other primary causal factors in *this* process (e.g. the precise orientation of the gun relative to Jones at the crucial moment) were not sufficient for the realization of the nominal essence of the death: if they had occurred without the first shot, and if certain events outside the process had failed to occur (e.g. the second shot), Jones would not have died at all. So the first

shot was a cause of Jones's death, as desired. But there was no causal process leading from the second shot to the death—i.e. to the realization of its real essence. If there had been no first shot, there would have been a process leading from the second shot to a death; but it would have been a slightly later death, a death with a different real essence. So the second shot was *not* a cause of Jones's death, as desired.

Objection: what do you mean, 'Jones would not have died at all [if there had been no first shot, and if certain events outside the actual process (e.g. the second shot) had failed to occur]'? Of course Jones would have died eventually. *Reply*: Jones would still have died, even if neither shot had occurred. But only because something else would then have caused his death. If *all* the potential causes of his death had failed to occur, he would have lived for ever.

We also had a more complicated version of the case, in which the noise of the second shot caused Jones to twitch so that the two bullets passed through a more vulnerable part of his body. Jones actually died, let us say, at noon. Without the second shot he would have died at 1 p.m. Without the first shot he would have died at a second past noon. But it was still the first shot which caused his death.

In this case *both* shots were factors in the actual process which led to Jones's death. The real essence of the death includes the fact that he died at noon precisely, and that would not have happened if either shot had been absent. But if there had been no second shot, and all the other primary causal factors in this process had occurred (including, for example, the first shot, the precise orientation of the gun at the time of the first shot, and so on), Jones still would have died. So the second shot was not a cause of his death. It affected his death (it hastened it), but did not cause it. Or at least it did not cause it *qua* 'his death'—'The second shot caused his death' is false, although 'The second shot caused his *early* death' may be true. On the other hand, the first shot *was* a necessary condition for his dying at all, by the process which actually led to his death. In its absence Jones would have died, certainly, but not by the process which actually led to his death. The precise orientation of the gun at the time of the *second* shot, for example, which played no part in the actual process, would then have become crucial. So the first shot did cause his death, as desired.

Now let us check that the analysis correctly distinguishes the two versions of Death in New York. In the Advice version, my action turns out to be a cause of Jones's death, as desired. There was a causal process leading from the advice to the death (at the later time t_2):

$$\left.\begin{matrix}\text{advice}\\ \text{Jones's mental state}\end{matrix}\right\} \longrightarrow \left.\begin{matrix}\text{goes to NY}\\ \text{violence in NY}\end{matrix}\right\} \longrightarrow \text{dies at } t_2$$

And the primary causal factors shown seem to constitute a minimal sufficient condition for his dying at all, relative to this process. The advice was not superfluous. Without the advice he would have died, certainly, by the following process:

$$\left.\begin{matrix}\text{no advice}\\ \text{Jones's mental state}\end{matrix}\right\} \longrightarrow \left.\begin{matrix}\text{does not go}\\ \text{tree-fall}\end{matrix}\right\} \longrightarrow \text{dies at } t_1$$

But this is a *different* process: it essentially involves the tree-fall, which played no part in the actual process.

The Massage version is different. There was a causal process *P* leading from the massage to his death (at the later time t_2):

$$\left.\begin{matrix}\text{massage}\\ \text{Jones's heart state}\end{matrix}\right\} \longrightarrow \left.\begin{matrix}\text{does not die at } t_2\\ \text{Jones's love of NY}\end{matrix}\right\} \longrightarrow \left.\begin{matrix}\text{goes to NY}\\ \text{violence in NY}\end{matrix}\right\} \longrightarrow \begin{matrix}\text{dies}\\ \text{at } t_2\end{matrix}$$

But here the primary causal factors shown were not a minimal sufficient condition for his dying at all, relative to this process. The massage *was superfluous. Without the massage he would still have died,* by the following process:

$$\left.\begin{matrix}\text{no massage}\\ \text{Jones's heart state}\end{matrix}\right\} \longrightarrow \text{dies at } t_1.$$

And this does not involve any primary causal factors beyond those of the actual process. The set {Jones's heart state, Jones's love of NY, violence in NY} was a sufficient condition for NO(Jones's death), relative to *P*. The massage did not cause his death, although it affected its time of occurrence (it delayed it).

Objection: we should distinguish those aspects of Jones's heart condition which were responsible for his not dying when given the massage from those which would have been responsible for his dying in the absence of the massage. Then it appears that the actual process was

$$\left.\begin{matrix}\text{massage}\\ \text{heart state 1}\end{matrix}\right\} \longrightarrow \left.\begin{matrix}\text{does not die at } t_1\\ \text{Jones's love of NY}\end{matrix}\right\} \longrightarrow \left.\begin{matrix}\text{goes to NY}\\ \text{violence in NY}\end{matrix}\right\} \longrightarrow \text{dies at } t_2$$

and the alternative process would have been

$$\left.\begin{matrix}\text{no massage}\\ \text{heart state 2}\end{matrix}\right\} \longrightarrow \text{dies at } t_1.$$

The alternative process *would* have involved primary causal factors beyond those of the actual process: the set {heart state 1, Jones's love of NY,

violence in NY} was *not* a sufficient condition for NO(Jones's death), relative to *P*—the massage needs to be included; so the massage did cause his death.

Reply: heart state 2 was not a necessary condition for his dying in the absence of the massage (relative to the actual process). If there had been no massage, and if heart state 2 had been absent, he would not have died at t_1, but he would have then gone to New York and died at t_2, *by the actual process*:

$$\left.\begin{matrix}\text{no massage}\\ \text{no heart state 2}\end{matrix}\right\} \longrightarrow \left.\begin{matrix}\text{does not die at } t_1\\ \text{Jones's love of NY}\end{matrix}\right\} \longrightarrow \left.\begin{matrix}\text{goes to NY}\\ \text{violence in NY}\end{matrix}\right\} \longrightarrow \begin{matrix}\text{dies}\\ \text{at } t_2\end{matrix}$$

Given the occurrence of the events in the set {heart state 1, Jones's love of NY, violence in NY}, Jones would have died whether or not he had got the massage, *and whether or not heart state 2 obtained*: so {heart state 1, Jones's love of NY, violence in NY} *was* sufficient for NO(his death), relative to the actual process.

Acknowledgement

I am indebted to an anonymous referee for several helpful comments.

Department of Traditional and Modern Philosophy
University of Sydney
Sydney 2006
Australia

References

Bunzl, M. [1979]: 'Causal Overdetermination', *Journal of Philosophy*, **76**, pp. 134–50.

Ehring, D. [1987]: 'Causal Relata', *Synthese*, **73**, pp. 319–28.

Hausman, D.M. [1992]: 'Thresholds, Transitivity, Overdetermination, and Events', *Analysis*, **52**, pp. 159–63.

Lewis, D. [1986]: *Philosophical Papers Vol. II*, Oxford, Oxford University Press.

McDermott, M. [1995]: 'Lewis on Causal Dependence', *Australasian Journal of Philosophy*, **73**, pp. 129–39.