CAUSATION AND THE PRICE OF TRANSITIVITY

That causation is, necessarily, a transitive relation on events seems to many a bedrock datum, one of the few indisputable a priori insights we have into the workings of the concept. Of course, state this *transitivity thesis* so boldly and it will likely come under dispute; one can reasonably worry that the appearance of an a priori insight glosses over a lack of cleverness on our part. Might not some ingenious counterexample to transitivity lurk nearby, waiting for a philosopher acute enough to spot it?

"Yes," comes the recent reply. In the last several years, a number of philosophers, myself included, have exhibited examples that appear to undermine transitivity (though, as we shall see, the target of my own examples was, and remains, quite different). I shall review a representative sample in section II. First, I need to sketch the main argument of the paper, and take care of some preliminaries.

I. INTRODUCTION AND PRELIMINARIES

I claim that the examples do have something to teach us about the metaphysics of causation, but that it is emphatically not that transitivity fails. Close inspection reveals that they pose no threat to that thesis. But they—or rather, certain of them, constructed with sufficient care—do show that transitivity conflicts with the following *dependence thesis*:

> Dependence: necessarily, when wholly distinct events c and e occur, and e counterfactually depends on c (read: if c had not happened, e would not have), then c is thereby a *cause* of e.

Dependence should not be confused with the transparently false claim that counterfactual dependence is necessary for causation; it is rather the claim that such dependence is sufficient—at least, when the events in question are wholly distinct, so that the relation of dependence does not hold merely in virtue of their mereological or logical relationships.[1] Nor is dependence the claim that causation can be given a counterfactual analysis, though most (if not all) counterfactual analyses endorse it. One might hold that no analysis of causation is possible, and yet endorse dependence. Or again, one might hold that it is counterfactuals that demand an analysis in terms of causation, and yet maintain dependence. What is more, dependence surely has a great deal of intuitive plausibility, at least

[1] For valuable discussion of this point, see David Lewis, "Events," in his *Philosophical Papers*, Volume II (New York: Oxford, 1986), pp. 241-69.

0022-362X/00/9704/198–222

when we are careful to read the counterfactual in a suitable "non-backtracking" sense, one which rules out such judgments as that, if *c* had not occurred, then it would have to have been the case that those earlier events which were causally sufficient for it did not occur.[2] Given the plausibility of dependence, its conflict with transitivity is all the more striking.

In the end, I shall suggest that were we forced to choose, we should reject dependence in favor of transitivity. But I shall also suggest that we are not forced to choose, since we can perfectly well recognize different *kinds* of causal relation in which events can stand, so that an obvious resolution of the conflict between dependence and transitivity presents itself: the kind of causal relation characterized by the first thesis is not the same as—indeed, fundamentally differs from—the kind of causal relation characterized by the second. Still, a manifest asymmetry distinguishes these causal relations, in that the kind of causal relation for which transitivity holds is clearly the more central of the two. So the price of transitivity—a price well worth paying—is to give up on the claim that there is any deep connection between counterfactual dependence and (the central kind of) causation.

Here is the plan for getting to these conclusions. First, I shall run through some of the alleged counterexamples to transitivity and highlight their apparent common structure (section II); examples with this structure I call *short circuits*. An immediate problem will emerge for my claim that dependence, and not transitivity, is the culprit; for while that claim neatly deflects some of the counterexamples (call these the *easy* ones), it does not help deflect the others (call these the *hard* ones). But the easy and hard cases share the short-circuit structure, and what is worse, it seems that it is in virtue of this structure that they are (or at least appear to be) counterexamples to transitivity. So targeting dependence seems quite mistaken.[3]

[2] See, for example, Lewis, "Causation," this JOURNAL, LXX, 17 (October 11, 1973): 556-67, reprinted with "Postscripts A-F to 'Causation'," in his *Philosophical Papers*, Volume II, pp. 159-213.

[3] An autobiographical note: it was precisely because of this problem for my thesis that transitivity conflicts with dependence that I became interested in these counterexamples in the first place. For it had seemed clear enough to me that there were cases—my easy cases—that exhibited this conflict; but then I found in the literature other cases, posed not as counterexamples to dependence but as counterexamples to transitivity, and I did not see how my thesis bore on them. Had I simply misdiagnosed my own cases, construing them as trouble for dependence when, in fact, that principle was quite irrelevant, and they were—like their cousins in the literature—counterexamples to transitivity, pure and simple? No, I had not. But it took me a while to see why not.

Not so, but it takes some work to see why. I start by sharpening the challenge the hard cases pose to transitivity by distinguishing three different problems for that thesis, arguing that they are not deep problems, and observing that their solutions help not at all in dealing with the hard cases, at least if those cases are formulated carefully enough (section III). The main argument follows: in section IV, I shall introduce a kind of causal structure which I call a *switch*, in which one event c interacts with a causal process in such a way as to redirect the causal route by which that process brings about a given event e. This structure differs in obvious ways from the apparent common structure of the alleged counterexamples. It also might seem to pose trouble of its own for transitivity, for reasons that I explain. I shall argue that the trouble is illusory, and that a number of considerations tell in favor of counting c a cause of e, when c is a switch with respect to e. That argument sets the stage for section V, where I revisit the alleged counterexamples to transitivity, and show that what distinguishes the easy cases from the hard ones is exactly that the hard ones, in addition to possessing the short-circuit structure, also possess a (much less noticeable) switching structure. So the intuitive verdict—that they provide us with cases in which c is a cause of d, and d of e, but not c of e—is mistaken: c is a cause of e, in virtue of being a switch with respect to e. Since the easy cases do not exhibit the switching structure, no such maneuver helps neutralize their threat to transitivity. But the original, obvious option—treat them as counterexamples only to dependence—remains available, and should be taken.

I close, in section VI, by highlighting two different lessons we learn from the examples. The first is metaphysical: what the easy cases show us is that transitivity and dependence conflict. Of course, that verdict leaves it open which one must go; I shall sketch a number of reasons to hold onto transitivity at the cost of dependence. The second is methodological: briefly, the hard cases remind us that the business of mining for intuitions about causation is risky: it is all too easy to mistake fool's gold for the real thing, and one must therefore be careful to subject one's hypothetical cases to careful analysis before trying to buy fancy conclusions with them. The easy cases stand up under such analysis; the hard cases do not.

II. THE COUNTEREXAMPLES AND THEIR APPARENT COMMON STRUCTURE

Here is an *example* from Michael McDermott[4]: a man plans to detonate a bomb. The day before, his dog bites off his right forefinger,

[4] In "Redundant Causation," *British Journal for Philosophy of Science*, XLVI (1995): 523-44.

so when he goes to press the button he uses his left forefinger instead. Since he is right-handed—and so would have used his right forefinger—the dog bite causes his pressing of the button with his left forefinger. This event, in turn, causes the subsequent explosion. But, intuitively, the dog bite does not cause the explosion.

Here is another *example*, adapted from one devised by Hartry Field[5]: an assassin plants a bomb under my desk (philosophy is a dangerous business, you know). I find it, and safely remove it. His planting the bomb causes my finding it, which, in turn, causes my continued survival. But, intuitively, his planting it does not cause my continued survival.

Next, an *example* from Igal Kvart[6]: a man's finger is severed in a factory accident. He is rushed to the hospital, where an expert surgeon reattaches the finger, doing such a splendid job that a year later, it functions as well as if the accident had never happened. The accident causes the surgery, which, in turn, causes the finger to be in a healthy state a year later. But, intuitively, the accident does not cause the finger to be in a healthy state a year later.

Finally, a refreshingly nonviolent *example* of my own: Billy and Suzy are friends. She is mischievous, and he is forever trying to keep her out of trouble. Billy sees Suzy about to throw a water balloon at her neighbor's dog. He runs to try to stop her, but trips over a tree root and so fails. Suzy, totally oblivious to him, throws the water balloon at the dog. It yelps. She gets in trouble. Billy's running toward Suzy causes him to trip, which, in turn, causes the dog to yelp (by dependence: for if he had not tripped, he would have stopped her from throwing and so the dog would not have yelped). But, intuitively, his running toward her does not cause the dog's yelp.

This last example is an easy case—easy because it is so obvious how to respond to it in a way that safeguards transitivity. After all, the only sense in which Billy's trip *causes* the dog's yelp is that it prevents something—Billy continuing to run toward Suzy, reaching her in time to stop her from throwing the balloon, and so on—which, had it happened, would have prevented the yelp. But no causal process connects the trip to the yelp (and not because the connection is an example of magical *action-at-a-distance*). So we are well within our rights to deny that, in this and similar cases of *double prevention*, the sort of dependence the yelp has on the trip is causal dependence—

[5] Personal communication.

[6] See his "Transitivity and Preemption of Causal Relevance," *Philosophical Studies*, LXIV (1991): 125-60.

and if so, the example poses no threat to transitivity. Or, more cautiously, we might allow that the trip is, in some sense, a cause of the yelp—but that it is not this nonstandard sense of cause we have in mind when we assert transitivity, but rather the ordinary, garden-variety sense.[7] Still, no threat to transitivity. The casualty, rather, is dependence: it is either false, or must be taken to be true only of a sort of nonstandard, second-class kind of causation.

But this response falls limp when it comes to the three other cases. Take Kvart's, for example: it is not merely that the healthy state of the finger counterfactually depends on the surgery; no, a perfectly ordinary causal process also connects the two. Likewise, a perfectly ordinary causal process connects the accident to the surgery. So how can we blame dependence, and not transitivity, for the counterintuitive result? Worse, it might seem that I have misdiagnosed my own easy case. For observe that all four examples have the following salient causal structure in common: an event c occurs, beginning (or combining with other events to begin) some process that threatens to prevent some later event e from occurring (call this process *threat*). But, as a sort of side effect, c also causes some event d that counteracts the threat (call this event *savior*). So c is a cause of savior, and savior—by virtue of counteracting threat—is a cause of e. But—or so it seems to many philosophers—c is not thereby a cause of e, and so transitivity fails. That diagnosis seems to apply equally to all four examples. So should we not favor it over a diagnosis that fits only the last of the examples?

No. The correct picture is a more complicated one, according to which the last example deserves exactly the diagnosis I gave it, and the first three hard cases deserve a quite different diagnosis. And the reason is that the hard cases have a quite different causal structure from the last, easy case.

Bringing out this different structure will take some care. As a preliminary, I shall contrast the foregoing alleged counterexamples to transitivity with three very different kinds of counterexample. We shall see that these other counterexamples are relatively benign—in that they leave ample room for a mildly qualified version of transitivity—but that they are also quite unlike the apparently more virulent strain exhibited above.

[7] This is the response I favor. For reasons, together with much more detailed discussion of other "double-prevention" cases, cf. my "Two Concepts of Causation" (manuscript). For discussion of related issues, cf. my "Non-locality on the Cheap?" (manuscript) and "The Intrinsic Character of Causation" (manuscript).

III. CONTRAST CLASS: BENIGN COUNTEREXAMPLES TO TRANSITIVITY

Take any event *e*—for example, my recent pressing of the letter '*t*' on my keyboard. Begin to trace its causes. Its immediate causes are relatively few, and near to hand: some signals in neurons in my brain, arm, and hand, the prior presence of that bit of the keyboard, and so on. As we trace back further, though, the causes become more numerous, scattered, and miscellaneous: they will likely include my birth, as well as the births of all those who were instrumental in the design and manufacture of this keyboard; if so, the number of more remotely ancestral births that count as causes of my simple act of typing becomes truly staggering. And we have not even considered those parts of *e*'s causal history which concern more than its merely human aspect. Go back far enough, and it may be that every event (or at least: every event that takes place at that stage of *e*'s backward light cone) is a cause of *e*.

One might balk at calling all such events *causes* of *e*. Surely, they are not as much causes of *e* as are its most proximate causes. Fair enough. The simplest way to accommodate this intuition is to say that causation comes in degrees, so that the more proximate a cause is to a given effect, the greater the degree to which it is a cause of that effect. Perhaps, then, what happens as we travel backward down *e*'s causal history is that the degree of causal influence of the events we find gradually diminishes; perhaps it eventually diminishes to the point where we must say that we no longer have causes of *e* at all. Such a picture of causation is harmless enough, though it raises the question of how to measure the attenuation of causal influence along a causal chain. Assuming that question settled, the needed modification to transitivity is obvious. Ideally, it will have the following form: when *c* is a cause of *d* to degree *x*, and *d* is a cause of *e* to degree *y*, then *c* is a cause of *e* to degree $f(x, y)$—where the function *f* returns values lower than, but close to, the lesser of *x* and *y*. Assuming some low threshold value α such that for $\beta < \alpha$, 'is a cause of *e* to degree β' no longer implies 'is a cause of *e*', transitivity will still hold for choices of *c*, *d*, and *e* where the values *x* and *y* lie well above the threshold. Or so we can hope the story would go.

At any rate, this issue, however interesting, has no bearing on our topic, as the examples discussed in the last section purported to exhibit failures of transitivity at early (backwardly speaking) stages of an event's causal history—too early to pin the blame on the sort of gradual attenuation of causal influence under discussion here.

Two other kinds of counterexample to transitivity focus on the possibility of disconnect between the parts of *d* that *c* causes and the

parts of *d* that cause *e*. It might be that *c* counts as a cause of *d* by virtue of causing one part of *d*, but that *d* counts as a cause of *e* by virtue of the causal action of a different part. If so, it need not follow that *c* is a cause of *e*.

What distinguishes the two kinds of counterexample is the notion of part at issue. Begin with the ordinary, mereological notion. A dance performance, for example, has as distinct parts the performances of each dancer. Let *d* be such a performance, and suppose that it has parts d_1 and d_2, consisting of the performances of Billy and Suzy, respectively. Among the causes of d_1 is, let us suppose, Billy's joining the dance troupe several months earlier. And among the effects of d_2 is, let us suppose, the appearance of a favorable review in the next morning's paper (for Suzy is the star of the show). We might count Billy's joining as a cause of the dance, in virtue of its role in causing one part of the dance (namely, Billy's). And we might count the dance as a cause of the favorable review, in virtue of the effect of one part of it (Suzy's brilliant performance). But we would not—or at least, not thereby—count Billy's joining as a cause of the favorable review. (For example, suppose he is a lousy dancer.) If so, we need to modify transitivity once again, presumably by introducing a distinction between primary and derivative senses of 'cause', so that (for example) it is only in the derivative sense that Billy's joining the dance troupe is a cause of the performance, and the performance of the favorable review, whereas it is only in the primary sense that causation is transitive. Again, interesting but irrelevant: for none of our counterexamples takes advantage of this loophole.

Finally, L. A. Paul[8] has highlighted a third—and arguably much more significant—way in which transitivity must be qualified. Put abstractly, we might have a situation in which *c* causes *d* to have a certain property or (to use Paul's terminology) *aspect A*, and *d*, in turn, causes *e*, in virtue of some other aspect *B*, but intuitively *c* is not a cause of *e*. Arguably, McDermott's "counterexample" provides a case in point: the dog bite does not cause the button pushing per se, but rather causes it to have a certain aspect—that is, causes it to be a "button pushing with the left hand." Moreover, the button pushing causes the explosion, but simply in virtue of being a button pushing; that it is a button pushing with the left hand is causally irrelevant to the explosion. Paul argues that in cases like this, transitivity—properly construed—simply does not apply: in order to apply, we must not merely have a single intermediate event to serve as a link between *c* and *e*, but must have a single intermediate event-cum-aspect.

[8] See her "Aspect Causation," this JOURNAL, this issue, pp. 235-56.

I think that is a very important point, and highly recommend Paul's expert treatment of it. But observe that there is no hope of applying it uniformly to cases with the short-circuiting structure. Consider, for example, Kvart's case: we cannot single out some aspect of the surgery, and argue that (i) the injury causes the surgery to have that aspect, but (ii) it is only in virtue of some other aspect that the surgery causes the finger to be healthy. Or again, a simple modification of McDermott's example renders it immune to Paul's treatment: suppose that after the dog bite, the man does not push the button himself but orders an underling to do so. The relevant intuitions do not change: the dog bite causes the order, and the order causes the explosion, but the dog bite does not cause the explosion. The only way I can see to apply Paul's observations is by way of a rather strained insistence that there is one event—call it a *making the button be depressed*—which the dog bite causes to have the aspect "being an order," and which otherwise would have had the aspect "being a button pushing." That resolution is unattractive enough to warrant the search for an alternative. So, while I certainly think that Paul's observation is very useful in a wide range of cases—indeed, I shall put it to use myself, in the next section—I do not think it provides the means for an adequate response to the challenge raised by the hard cases. In fact, I think the right response has a quite different shape: transitivity does apply to those cases, and yields the correct conclusion—intuition notwithstanding—that c is a cause of e. The next section provides the crucial underpinnings for this response.

IV. SWITCHES

A distinctive relationship that an event c can bear to an event e is that of helping to determine the causal route by which e is produced. In such cases, I call c a *switch* with respect to e. Here is an example:

> "The Engineer": an engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the right-hand track, instead of the left. Since the tracks reconverge up ahead, the train arrives at its destination all the same; let us further suppose that the time and manner of its arrival are exactly as they would have been, had she not flipped the switch.

Let c be the engineer's action and e the train's arrival. Pick an event d that is part of the train's journey down the right-hand track. Clearly, c is a cause of d and d of e; but is c a cause of e? Is her flipping the switch a cause of the train's arrival? Yes, it is, though the opposing reaction surely tempts. "After all," goes this reaction, "is it not clear that the switching event makes no difference to whether the train arrives, but merely determines the route by which it arrives?"

To some extent we can accommodate this reaction: yes, of course, the switching makes no difference to the arrival, in the sense that had it not occurred, the train would have arrived all the same. And yes, of course, the switching helps determine the causal route to the arrival. But it goes too far to conclude that the switching is not itself among the causes of the arrival. A number of considerations reinforce this conclusion; since the issue is crucial to the proper diagnosis of the hard cases, it will pay to review these considerations with some care.

The explanation test. Ask for a complete explanation of the train's arrival—"How did it get there? What processes led to its arrival?"—and a complete answer must include the switching event. But, plausibly, this complete answer will simply consist in a description of the arrival's causal history.

The subtraction test. Remember that the train travels down the right-hand track. Certain rather boring events extraneous to this causal process are the persistence through time of the various bits of left-hand track; these are events which would have combined with the presence of the train to cause its arrival, had it traveled down the left-hand track. Suppose some of these events away; that is, subtract them (you need not go so far as to replace them by pure void; air and dirt will do). Subtract enough of them, in fact, that the left-hand track ends up having a sizable gap in it. Plausibly, this subtraction should not alter the facts about the arrival's causes, since the events subtracted are quite separate from and extrinsic to the events that constitute the causal processes leading up to the arrival.[9] But in the altered situation, the engineer's action quite obviously helps bring about the train's arrival; after all, she steers the train onto the only track that will get it to its destination. So, given that subtracting extraneous events should not alter the causal history of the arrival, we should also say that in the original case, her switching is part of this causal history.

More generally, when a switch alters the causal pathway of some process *M*, there will typically be other processes with which *M* would have interacted to bring about the effect *e*, but with which—thanks to the switch—it does not interact. Since the events that go into these other processes are, as things stand, causally idle features of the environment, it seems plausible that we ought to be able to remove them from the environment without altering the causal status of the switch *c* with respect to the effect *e*. But in the scenarios where they are absent, it becomes perfectly clear that *c* is a cause of *e*.

[9] For detailed discussion and defense of the "intrinsicness" thesis being relied on here, cf. my "The Intrinsic Character of Causation."

The variation test. The example had a little intuition pump built into it, given that we stipulated that, if the engineer had not thrown the switch, the train would have arrived at exactly the same time and in exactly the same manner. Suppose we had told the story slightly differently, so that, if the engineer had not thrown the switch, the time and/or manner of the arrival would have been different—say, the train would have arrived an hour later. This alteration reverses, I think, the intuitive verdict (such as it was) that the switching is not a cause of the arrival. But that verdict should not be so sensitive to the details of the case—which, in turn, suggests that the no-difference-whatsoever intuition pump was misleading us.

Paul's observation. If the details of the train's counterfactual journey down the left-hand track are sufficiently similar to the details of its actual journey down the right-hand track—and it is certainly natural to interpret the story as if this were true—then the right thing to say might be that, at a suitably abstract (that is, nondetailed) level of description, the two journeys consist of exactly the same events. To borrow Paul's terminology, what the switching does is to make it the case that these events have certain aspects: specifically, the switching makes it the case that they are travelings-down-the-right-hand-track rather than travelings-down-the-left-hand-track. On such an analysis, no two-step chain of the right sort connects the switching to the arrival, and so transitivity simply does not apply. Perhaps our tendency to think of the actual and counterfactual journeys as really the same journey, differently realized, lies behind whatever reluctance we might feel to calling the switching a cause of the arrival.

As a test, hold the details of the arrival fixed, but alter the extraneous events so drastically that the way the train gets to its destination, in the counterfactual situation in which it travels down the left-hand track, is completely different from the way it, in fact, gets to its destination: it stops after a short while, gets taken apart, shipped piecemeal to a point near its destination, reassembled, and all this in such a way as to guarantee that nothing distinguishes its counterfactual from its actual arrival. Then Paul's observation does not apply, since there is no event that can happen as a traveling-down-the-right-hand-track, and as a shipping-of-disconnected-parts. But, by the same token, there is much less temptation to deny that the switching is a cause of the arrival.

Salience. Notice, first, that among the (highly nonsalient) causes of the train's arrival is the presence of a certain section *S* of the track down which the train travels, during the time that the train is, in fact, on *S*. And among the causes of the presence of *S* at this time is the

presence of *S* a day earlier. So we should say: among the causes of the arrival is the presence of *S* a day earlier. That is certainly odd, but it is odd in the familiar way to which writers on causation have grown accustomed: whatever the exact nature of the causal concept we are trying to analyze, it surely picks out a very permissive relation, one which does not distinguish between events that we would normally single out as causes and events we would normally ignore because their causal relationship to the effect in question is too boring or obvious to be worth mentioning, or is easily hidden from view as part of the "background conditions." The presence of *S* a day earlier is like that.

But the relationship that the switching event bears to the arrival is just like the relationship that the earlier presence of *S* bears to the arrival: for the relationship that the switching event bears to the setting of the switch as the train passes over it is just like the relationship that the earlier presence of *S* bears to the later presence of *S*, and the relationship the setting of the switch as the train passes over it bears to the arrival is just like the relationship the later presence of *S* bears to the arrival.

Switching per se is not causing. The last paragraph helps bring out something rather important, which is that the switching bears two distinct causal relationships to the arrival. The first is the highly nonsalient one just outlined: the switching event causes the setting of the switch, which interacts with the passing train in the same way as the mere presence of a piece of track interacts with the passing train, and so on. The second is the narratively more vivid relation: the engineer's action makes a difference to the causal route by which the arrival happens. But it is only in virtue of the first relation that the switching is a *cause* of the arrival. To see this, consider how the two relations can come apart, as in the following *example*: Billy and Suzy are about to throw rocks at a bottle. Suzy, the quicker of the two, is winding up when suddenly her muscle cramps, stopping her from throwing. So Billy's rock gets there first. Had there been no cramp, Suzy's rock would have struck the bottle before Billy's, breaking it—in which case Billy's rock would have struck only empty air. Either way, the bottle breaks; what the cramp does is to determine the causal route to the breaking (that is, whether it happens via Suzy's throw, or via Billy's). But the cramp is manifestly not a cause of the breaking. What this shows is that switching per se is not causing; that is, it does not follow from the fact that an event *c* determines the causal route to an event *e* that *c* is among *e*'s causes.

But recall that the *kind* of switch on which I am focusing—a kind of which the engineer's action is an instance—is an event *c* that in-

teracts with a causal process in such a way as to redirect the causal route by which that very process brings about a given event *e*. I claim that in cases of this kind of switching—call them cases of *interactive switching*—*c* is a cause of *e*. But that will be so in virtue of *c*'s interaction with the causal process that results in *e*, and not merely because *c* is a switch with respect to *e*. And this fact, in turn, helps account for our reluctance (such as it is) to count the switching as a cause of the arrival: we naturally focus on the narratively most salient relationship that the first event bears to the second—namely, that of helping to determine the causal route—and, recognizing that this relationship does not suffice for causation, conclude that there is none to be had.

An obvious test of this hypothesis is to construct a case where the *causal* role is the vivid one, and the switching role is not. Here is such a case. "The Kiss": Billy and Suzy have grown up. One day, they meet for coffee. Instead of greeting Billy with her usual formal handshake, however, Suzy embraces him and kisses him passionately, confessing that she is in love with him. Billy is thrilled—for he has long been secretly in love with Suzy, as well. Much later, as he is giddily walking home, he whistles a certain tune. What would have happened had she not kissed him? Well, they would have had their usual pleasant coffee together, and afterward Billy would have taken care of various errands, and it just so happens that in one of the stores he would have visited, he would have heard that very tune, and it would have stuck in his head, and consequently he would have whistled it on his way home, and, just to give the case the right "shape," let us stipulate that matters are rigged in such a way that the time and manner of the whistling would not have differed at all. So the kiss is an event that interacts with a certain process—call this process *Billy's day*—in such a way as to redirect the causal route by which that process brings about a certain effect (the whistling). But even though there is the failure of counterfactual dependence typical of switching cases (if Suzy had not kissed Billy, he still would have whistled), there is, of course, no question whatsoever that as things stand, the kiss is among the causes of the whistling.

Those who still balk at calling the engineer's action a cause of the train's arrival should answer the following question: What is the relevant difference between that case and this one? None, so far as I can see—that is, none relevant to the question whether the switching event *c* is a cause of the effect *e*. But there are clear and significant differences relevant to an explanation of why our intuitions about

"The Engineer" are less firm and settled than our intuitions about "The Kiss." For in "The Engineer," it is the switching role of the engineer's action that stands out, and not her causal contribution to the arrival; and switching per se is not causing. But in "The Kiss," the switching role is quite obscured—in fact, in telling the story one must forcibly draw attention to it—whereas the causal contribution of Suzy's kiss could not be more obvious.

To wrap up: the points canvassed in this section (i) provide us with ample positive reason to count interactive switches as causes; and (ii) provide us with several plausible explanations of the source of any contrary intuitions. I conclude that the price of accepting interactive switches as causes is quite small, certainly small enough to be worth paying. Especially so since, as we are about to see, doing so provides us the means to disarm the hard counterexamples to transitivity.

V. TWO KINDS OF SHORT CIRCUITS

Recall the apparent common structure of the counterexamples, common to both easy and hard cases: one event *c* helps to initiate a threat, which, if not stopped, will prevent *e* from occurring. But *c* also causes a savior event that, in turn, counteracts threat. So *c* causes savior, and savior—by counteracting threat—causes *e*. But *c* does not thereby cause *e*. Or so the stories go.

Now, however, we are in a position to see that there are two very different ways in which threat might be counteracted. I shall bring them out by means of examples, building on the case discussed in the last section.

> "Drunken Bad Guys": the bad guys want to stop the good guys' train from getting to its destination. Knowing that the switch is currently set to send the train down the left-hand track, the bad guy leader sends a demolition team to blow up a section of that track. En route to its mission, the team stops in a pub. One pint leads to another, and hours later the team members have all passed out. The good guys, completely oblivious to what has been going on, send the train down the left-hand track. It arrives at its destination.

> "Clever Good Guys": the bad guys want to stop the train from getting to its destination. Knowing that the switch is currently set to send the train down the left-hand track, the bad guy leader sends a demolition team to blow up a section of that track. This time, it is a crack team, not so easily distracted. The good guys have gotten wind of the bad guys' plans, however, and send word to the engineer to flip the switch. She does so. The bad guy demolition team blows up a section of the left-hand track,

but to no avail: thanks to the engineer's action, the train travels down the right-hand track. It arrives at its destination.

Let c be the bad guy leader's sending the team on its mission. Let e be the train's arrival. Let d_1 be the event, in "Drunken Bad Guys," of the team's stopping in the pub. Let d_2 be the event, in "Clever Good Guys," of the engineer's flipping the switch. d_1 and d_2 are, in each case, the savior events.[10] Clearly, c, in each case, causes them. But it turns out to be entirely too quick to assume that they have the same causal status with respect to e, for there is a world of difference between the ways that each event counteracts threat.

In "Drunken Bad Guys," what the savior event does is to cut short the threat process. Notice, furthermore, that the way that threat is cut short requires no causal connection whatsoever between the savior event (the team's stopping in the pub) and the final effect (the train's arrival). The intuition that c—the bad guy leader's sending out the team—is not a cause of e—the train's arrival—is quite correct; but that fact poses no threat to transitivity, for the intermediate event d_1 is likewise not a cause of e. Dependence may say otherwise—but so much the worse for that thesis.

The way the threat is counteracted in "Clever Good Guys," by contrast, is entirely different. The engineer's flipping of the switch does not cut short threat at all; on the contrary, that process goes to completion, resulting in the destruction of a portion of the left-hand track. Rather, what the savior event does is to switch the causal process leading to the train's arrival from a pathway that is vulnerable to threat to one that is immune to threat. Quite obviously, such a method for counteracting threat must involve a switch, of the sort examined in the last section. Therefore, given the conclusions of that section, this way of counteracting threat necessarily establishes a causal connection between the savior event and the final effect (assuming, as I shall throughout this discussion, that the switch is an interactive switch).

What of the event c, the bad guy leader's sending out the team? That too, in "Clever Good Guys," is a cause of the arrival, by virtue of being a cause of the engineer's flipping the switch. But is that not counterintuitive? Counterintuitive enough that we should reject the tacit appeal to transitivity made in the foregoing 'by virtue of' clause? Well, yes, it is counterintuitive—but there is excellent reason to think that intuition is being misled here. For what intuition naturally focuses upon is the narratively most vivid role that c plays in the story. And this is exactly the same "short-circuiting" role that c plays

[10] Of course, we could have chosen other events to play this role; nothing hinges on the particular choice made.

in "Drunken Bad Guys": it initiates a threat to *e*, and also causes that threat to be counteracted. Intuition is perfectly right to insist that no event *c* can be a cause of an event *e* merely by virtue of (i) doing something that threatens to prevent *e*; and (ii) doing something that counteracts that very threat. Indeed, that is the lesson of "Drunken Bad Guys," for in that story the only causal relationship *c* has to *e* is that of threatening to prevent *e* and simultaneously causing that threat to be counteracted. Intuition stridently objects to calling *c* a cause of *e*, in that case; and what is more, no argument can be found to persuade her to change her mind. So she should be heeded.

But in "Clever Good Guys," such an argument is available. For what we naturally fail to notice—given the narrative prominence of the short-circuiting role—is that *c* bears another causal relation to *e*, namely, *c* begins a process that ultimately interacts with the processes leading up to *e* in such a way as to alter the causal route they follow to *e*. All the reasons for counting switches as causes apply here with equal force. So we must patiently explain to intuition that she is being distracted by irrelevant details of the case, and should reverse her verdict.

One way to see that these details—specifically, the fact that *c* is a short circuit with respect to *e*—really are irrelevant is to compare "Clever Good Guys" to the following case:

> "Mischievous Bad Guys": this time, the bad guy leader has no intention of stopping the train from getting to its destination. But he knows the good guys are paranoid, and likes to play tricks on them. Just for fun, he sends out the demolition team—not with orders to blow up the left-hand track, but with orders to go down to the pub and enjoy themselves. Still, he knows full well that once the good guy spies have reported that the team has been sent out, the good guys will panic and order the engineer to flip the switch. Sure enough, they do. The engineer flips the switch. The train travels down the right-hand track. It arrives at its destination.

There is no short circuit here, because there is no threat to be counteracted. So in that sense, "Mischievous Bad Guys" differs dramatically from "Clever Good Guys." But in the important sense, the two cases are exactly the same: for the relevant relationship the bad guy leader's action bears to the arrival differs not at all. What is more, there is no question (not, at least, once we absorb the lessons of the last section) that in "Mischievous Bad Guys," the leader's action is a cause of the arrival—a verdict we should therefore carry over to "Clever Good Guys." The fact that in this latter case, the leader's action is also a short circuit with respect to the arrival mat-

ters not one whit; rather, it serves merely as a distraction, making the case difficult to judge, and off-hand intuitions about it quite suspect.

This diagnosis neatly extends to the hard cases introduced in section II. Consider Kvart's (I shall leave the others as an exercise). Does the injury cause the finger to be healthy? Intuitively no, but on inspection yes: for the accident that befalls the man redirects a certain process (namely, the process consisting, roughly, of him and his movements) onto a causal pathway different from the one it would have followed, and, in fact, both pathways (actual and counterfactual) issue in the same result (a healthy finger). So the accident is a switch with respect to the healthy finger. Observe, furthermore, how easily we can apply the various considerations canvassed in the last section to reinforce this conclusion. I shall consider just the variation test: alter events in the environment sufficiently to make it the case that, if the accident had not occurred, the finger's state of health would have been worse (say, because the man would have suffered some other, much more serious accident). Then we would not hesitate to count the accident as one of the causes of the finger's later state of health.[11] Again, what this test brings out is the (somewhat hidden) switching structure of the example; that it has the short-circuiting structure as well is merely a distraction.

Here, then, is the point we have come to: (i) when c is a switch with respect to e (of the interactive kind), then c is a cause of e, although for various reasons this fact may not be immediately obvious. (ii) What distinguishes the hard cases from the easy ones is exactly that, whereas both exhibit events c and e such that c is a short circuit with respect to e, in hard cases it is also true that c is a switch with respect to e. (iii) When we render the intuitive verdict about hard cases that c is not a cause of e, we wrongly focus solely on the fact that c is a short circuit with respect to e. Observing (correctly enough) that c is not thereby a cause of e, we wrongly conclude that it is not a cause of e, simpliciter. But c is also a switch with respect to e, so that conclusion is mistaken. The mistake is natural enough, though, given that the short-circuiting role is by far the more vivid one. (iv) Since, however, no such diagnosis is available for the easy cases, the intuitive verdict that, in such cases, c is not a cause of e stands. But in such cases, we can always find an intermediate event d such that c is

[11] A slightly different consideration pointed out to me by Tim Maudlin also applies: alter the surgery enough so that the finger ends up in a much better state of health than it would have without the accident. Here, too, we would not hesitate to count the accident as a cause of the finger's state of health. The relevant methodological point still applies: prima facie, the correct judgments about the causal status of the accident should not be so sensitive to these fine details.

clearly a cause of d, and e counterfactually depends on d, but does so in the odd double-preventing kind of way. Hence such cases exhibit an intractable conflict between dependence and transitivity. (v) Since, in such cases, there are independent grounds for doubting that d is a cause of e (most notably: no causal process connects them), the appropriate response is to deny dependence.

Points (iv) and (v) deserve more discussion; I shall reserve discussion of (v) for the next section. Now, given that I have been at such pains to find grounds for resisting the intuitive verdict about the hard cases, one might reasonably wonder whether, with enough ingenuity, we might find similar grounds for resisting the intuitive verdict about the easy cases. I say "no": the intuitive verdict about those cases is correct, and that is why they have something useful to teach us about causation (namely, that counterfactual dependence does not suffice for it). While I cannot hope to establish this claim conclusively, I can at least show that not one of the strategies that helped us with the hard cases helps in the slightest with the easy cases. Let us review them, using "Drunken Bad Guys" as our canonical example of an easy case.

The explanation test: Does the fact that the bad guy leader sent out the demolition team help explain the train's arrival? Of course not. The closest this fact can get to appearing in an explanation of the arrival is to appear in a very specific sort of explanation request, namely, we can ask: "Why did the train arrive at its destination, given that the bad guy leader sent out a demolition team to stop it from doing so?" The obvious and correct reply is to cite the relevant double preventer: the demolition team got side-tracked by the pub; this event prevented it from doing something (blowing up the track) which, in turn, would have prevented the train's arrival. Keeping in mind that the 'given that' clause might well be tacitly assumed in the context of the explanation request, it seems plausible that double preventers of the sort exhibited by this example can only be explanatory relative to this specific sort of explanation request. But in the context of such a request, it would be silly to add the content of the 'given that' clause as an extra bit of explanatory information. No: on the assumption that the 'given that' clause cites a genuine threat (which we can take to be a presupposition of the why-question), the question is answered completely by citing those events which counteracted the threat, and explaining how they did so.

The subtraction test: intuitively, the bad guy leader does not help cause the train's arrival by sending out the demolition team. Can we reverse this intuitive verdict by *subtracting* events from the environment of the processes that issue in the arrival? At the very least, can a

selective subtraction make it the case that the arrival counterfactually depends on the leader's action? Well, let us review the events that are available for subtraction. On the one hand, there are the events that make up the train's journey down the tracks, and the rather less noticeable events that consist in the persistence of the appropriate bits of track. No doubt there are other even less noticeable events that contribute to the arrival: the presence of sufficient oxygen to allow the train's engines to burn fuel, and so on. We need not canvass them all, for it is clear that the most that can happen if we subtract some of these events is that the arrival will not happen. And that is, of course, not a situation in which (i) the arrival happens; and (ii) its occurrence counterfactually depends on the leader's action.

Where else to look? Well, there are the events that make up the rest of the story: the leader's action itself, the team's journey, the drunken revelry in the pub, and so on. Subtract some of these, and the most that will happen is that we get a situation in which the team succeeds in stopping the train (for example, if we remove the pub itself). No help there, either. So the example fails the subtraction test.

The variation test: it likewise fails this test. Remember that the object here is to *alter* events selectively so that the manner of the arrival at least depends on the leader's action. Keeping in mind the different sorts of events that are available, it becomes clear at once that the only effect such alterations could possibly have is to make the leader's action *prevent* the arrival (for example, alter the potency of the beer in the pub sufficiently so that the team does not become drunk, and therefore continues on its mission). So the example fails the variation test, as well.

It was clear all along that it was bound to fail both the subtraction test and the variation test. For what those tests bring out, when they succeed, is that (i) there is an alternative causal pathway to the given effect e; (ii) the effect of the alleged cause c is to switch the processes leading to e away from this alternative and onto a different path; and therefore (iii) altering the character of the path not chosen—either in an extreme way (the subtraction test), or in a modest way (the variation test)—will yield a situation in which e, or the manner of e's occurrence, does depend on c. For, in this new situation, if c had not happened, then the processes aimed at e would have followed the alternative (and now altered) causal pathway, and therefore either would have missed their mark (the subtraction test), or would have produced e in a slightly different manner (the variation test). But in "Drunken Bad Guys"—as, indeed, in all easy cases—the al-

leged cause simply does not act as a switch. So, of course, the subtraction and variation tests fail, for such cases.

Paul's observation: Could it be that transitivity simply does not apply, because for any candidate intermediate event d, we shall find that while in a loose sense, c (the leader's action) causes d and d causes e (the train's arrival), what happens, strictly speaking, is that c causes d to have a certain aspect, but it is only in virtue of some other aspect that d causes e? Let us find out, by taking our intermediate d to be the team's decision to stop in the pub. Now, dependence (and common sense) yields the verdict that the leader's action is a cause of d; likewise, dependence (but not, this time, common sense) yields the verdict that d is a cause of the arrival. But as far as I can tell, *aspect* intuitions are silent: it is not that we can fix on some way that d happens, and say with confidence that what the leader's action really does is to cause d to happen in this way; likewise for the relationship d bears to e. So this strategy is not available: the defender of dependence cannot say (not on these grounds, anyway) that transitivity does not apply, and so the easy cases exhibit no conflict between that thesis and dependence.

Salience: return to the case of "The Engineer." In that case, it can be agreed on all sides that the engineer's action causes an event—the setting of the switch as the train passes over it—that interacts with the processes that lead to the arrival; the question was whether this interaction is of the right sort to qualify the engineer's action as one of the arrival's causes. And we were able to answer that question in the affirmative by observing that the setting of the switch bears the same sort of relationship to the arrival as other events that clearly are causes of the arrival, albeit highly nonsalient ones (for example, the presence of the bits of track over which the train passes). But if we try to apply this strategy to the case at hand, arguing that the bad guy leader's action is simply a highly nonsalient cause of the arrival, we cannot even get it off the ground—for the leader's action causes no event that interacts at all with the processes that lead up to the arrival, let alone interacts with them in the right sort of way.

To put the point rather mildly, we appear to have run out of options for defending the claim that the leader's action is also a cause of the arrival, and intuition be damned. I conclude that we should side with intuition in this, as in all other easy cases. The conflict between dependence and transitivity is unavoidable.

VI. METAPHYSICAL AND METHODOLOGICAL LESSONS

Which to give up? Pose the question abstractly, and it may be hard to decide. But, in fact, the details of the examples that exhibit the

conflict make the question easy: for what is so striking about those examples is that the intermediate event *d causes* the effect *e* only in the sense that it prevents something that would have prevented *e*. Thus, Billy's tripping stops him from preventing Suzy's throw; the demolition team's decision to stop in the pub winds up preventing them from blowing up the track, which would, in turn, have prevented the train's arrival; and so on. No causal process connects *d* with *e*; *d* does not interact with other events to help bring about *e*; and so on. Of all of the characteristics we expect from the causal relationship, the only one exhibited by the relationship between *d* and *e* is that *e* counterfactually depends on *d*. And even that characteristic is entirely optional, as witness standard cases of preemption, in which one event causes a second, even though an alternative, preempted process would have brought about the second, had the first not done so. The striking conflict with transitivity thus confirms a suspicion that should have been there from the outset: namely, that double prevention is not causation. If we had independent reason to think that transitivity fails, then we might hold on to dependence, even in the face of the examples. But once the hard cases have been defused, we have no such independent reason.

We do have independent reasons to doubt dependence, on the other hand—reasons that go beyond (or, perhaps, simply properly articulate) the doubts about that thesis which naturally arise when we focus on cases of double prevention. For example, dependence can be shown to conflict both with a prohibition on action-at-a-distance, and with the claim that the causal structure of a process is determined by its intrinsic, noncausal character, together with the laws that govern it. But since that story is long—and I have told it elsewhere[12]—I shall not go into detail here. Given my focus on transitivity, however, it is appropriate to point out two new sorts of trouble that arise if we extend dependence in a natural way to cover cases of prevention, and of causation by omission. More exactly, suppose we endorse the following theses:

(1) Event *c* causes the omission of an event of type *E* if it is the case that *c* occurs, no event of type *E* occurs, and if *c* had not occurred, an event of type *E* would have occurred.

(2) The omission of an event of type *C* causes event *e* if it is the case that no event of type *C* occurs, *e* occurs, and if an event of type *C* had occurred, then *e* would not have occurred.

[12] Cf. my "Two Concepts of Causation."

Thus, (1) is a natural application of dependence to causation of omission, and (2) is a natural application of dependence to causation by omission.

Then, if transitivity holds, we get wonderful results, as in the following three cases:

> *Preemptive strike:* one day, Suzy suddenly and viciously attacks Billy—so viciously that he is quite incapacitated, as a result. Asked to explain herself, she states that she was merely preventing him from attacking her. "But," her interrogators exclaim, "you know full well that he had no intention of attacking you—the two of you were friends!!" "Yes," she replies, "he had no intention of attacking me then. But he certainly does now—so it's a good thing I incapacitated him."

Unfortunately, the combination of dependence (extended as in (1)) and transitivity endorses this reasoning. For her attack causes his subsequent incapacitation; and, were he not incapacitated, he would attack her, whereby his incapacitation causes his failure to attack her. By transitivity, her attack did indeed cause him to fail to attack her.

> *No cure:* Billy is sick, and needs a certain drug to cure his disease. But the cure requires two doses, whereas only one is available. (Suppose further that one, by itself, has no effect on the disease.) In a gesture of futility, Dr. Jones gives Billy the one dose on Monday. On Tuesday, Dr. Smith comes in, intending to do the same thing, but of course failing to. Billy remains sick. What is worse, Dr. Jones gets blamed for Billy's continued ill health, on the following grounds: by giving Billy the dose on Monday, he caused Dr. Smith's failure to give Billy the dose on Tuesday (for if Jones had not given the dose, Smith would have). But Smith's failure to give the dose on Tuesday, in turn, caused Billy to remain ill, since if Smith had given Billy the dose, then—with one dose already in his system—Billy would have been cured. By transitivity, Jones made Billy remain sick.

> *Collision avoidance:* the engineer flips the switch, sending the train down the right-hand track; recall that the tracks reconverge up ahead. She promptly puts in for merit pay, on the grounds that she has prevented a collision. After all, she causes the absence of a train on the left-hand track (for if she had not flipped the switch, the train would have been there), which, in turn, causes there to be no collision (for if there had been a train on the left-hand track, there would have been a collision at the point of reconvergence). So her action caused the omission of a collision.

Now, it is certainly true that in all of these cases, there is a great temptation to deny the relevant counterfactuals. For example, we all

surely want to insist that, if there had been a train on the left-hand track, that would have been because the engineer did not flip the switch, and so there (still) would have been no collision. But however sensible such *backtracking* reasoning is, it is simply not available to the loyal defender of dependence, since that thesis requires a scrupulously nonbacktracking reading of the counterfactual. So the loyal defender must, on pain of denying transitivity, embrace these remarkable results. All the more reason to shift allegiance elsewhere.

Moreover, the cost of denying dependence should not be overestimated. We certainly need not abandon the truism that where there is counterfactual dependence between distinct events, there is typically causation. Nor, I think, need we deny that such dependence is a kind of causal relation—so long as we are clear that it is not the kind of causal relation we have in mind when we think of processes, interaction, and transitivity. For obvious reasons, we might call this latter kind of causal relation *production*; our thesis should then be that production and counterfactual dependence are both causal relations, but that only production obeys transitivity.[13] To this claim we should add that counterfactual dependence is not the central kind of causal relation. As evidence, observe that there is a clear asymmetry in our intuitive reactions to cases of counterfactual dependence without production, on the one hand, and cases of production without counterfactual dependence, on the other. As an example of the latter, suppose Billy and Suzy both throw perfectly aimed rocks at a window. Suzy's gets there first, breaking it. Intuition unhesitatingly picks out her throw, and not Billy's, as a cause of the breaking. But when we have a case of dependence without production—as in "Drunken Bad Guys," for example—intuition is more equivocal; it makes sense, for example, to ask whether, by entering the pub, the demolition team really helps cause the train's arrival. Why such a question makes sense is perfectly clear: by stressing the word 'really' (or functional equivalents such as 'strictly speaking', and the like), one invokes a context where the central kind of causation is salient; and entering the pub is not a cause of the arrival, in this sense.

Let us now turn from the metaphysical to the methodological: What lessons does our discussion have to teach us about how we ought to conduct a philosophical investigation of causation? Very simply put, the lesson is this: treat intuitions about cases with care. It will not do simply to construct hypothetical cases, consult our off-

[13] Certain other principles of interest distinguish the two varieties of causation as well; cf. my "Two Concepts of Causation."

hand intuitions about them, and without further ado use the results to draw sweeping (or less-than-sweeping, for that matter) conclusions about the nature of causation. For such an approach offers no insight into the source of our intuitions, and understanding that is crucial if we are to know what, if anything, these intuitions have to teach us. To gain such insight, there is no substitute for close inspection of the cases—close enough that we can be confident that we have adequately discerned their salient structure.

Ignoring this point can lead to either of two opposing errors, both of which are to be found in the literature. The first consists in excessive gullibility, a willingness to take intuitions about cases at face value, without interrogating their credentials. Thus McDermott, when discussing his dog-bite case as well as others, fails to conduct anything like a detailed analysis of his cases, and instead rests content with reporting the results of various informal polls (of "naive subjects," no less) about them—as if these results could establish anything more than the unsurprising result that most of us find it counterintuitive to count the dog bite as a cause of the explosion. Of course, that is counterintuitive, and, of course such counterintuitive results count as prima facie reason to doubt transitivity—but prima facie reasons can wither upon scrutiny, and these do. McDermott, unfortunately, seems not to see the need for such scrutiny.

David Lewis,[14] in his discussion in this issue, makes a rather more interesting mistake. Notice first the way he characterizes the alleged counterexamples to transitivity:

> The counterexamples have a common structure. Imagine a conflict between Black and Red. (It may be a conflict between human adversaries, or between nations, or between gods striving for one or another outcome, or just between those forces of nature that conduce to one outcome versus those that conduce to another.) Black makes a move that, if not countered, would have advanced his cause. Red responds with an effective countermove, which gives Red the victory. Black's move causes Red's countermove, Red's countermove causes Red's victory. But does Black's move cause Red's victory? Sometimes it seems not (*ibid.*, p. 194).

Now, one can argue about whether Lewis has adequately captured the common structure of the counterexamples. (What is Red's countermove in "Drunken Bad Guys"? Putting a pub in the path of the team?) Never mind: what is important to notice is that Lewis

[14] Lewis, "Causation as Influence," this JOURNAL, this issue, pp. 182-97.

completely glosses over the crucial distinction between counterexamples that do and counterexamples that do not involve switches. So, when he goes on to say that his "considered opinion is that Black's move does indeed cause Red's victory" (*ibid.*), one must ask why this considered opinion places no importance on the switching/nonswitching distinction.

A possible answer comes from examining Lewis's defense of his position. While noting that it is somewhat counterintuitive, he goes on to insist: "Insofar as I can summon up any inclination to accept the counterexamples, I think my inclination has three sources, all of them misguided" (*ibid.*). He observes that (i) in many of the cases, "Black's move prevents Red's victory as well as causing it: it causes one version, but it prevents another"; (ii) "Moves such as Black's are in general conducive to victory for Black, not for Red"; and (iii) Red's victory does not counterfactually depend on Black's move (*ibid.*, pp. 194-95). Since none of these bad reasons for refusing to call Black's move a cause of Red's victory turns on the switching/nonswitching distinction, perhaps that is why Lewis ignores it.

That is a mistake. To begin with, Lewis has not looked far enough for sources of the "inclination" he wishes to resist. The best way to see this is to observe how explanatorily inadequate are the sources he cites: (i) it is child's play to come up with counterexamples in which Black's move makes no difference whatsoever to the manner of Red's victory, and so cannot be said to prevent one version of it while causing another. ("Drunken Bad Guys" is such a case.) (ii) It is likewise child's play to come up with counterexamples in which Black's move threatens Red's victory only thanks to the existence of exceedingly improbable circumstances, and so cannot be said to be of a type "in general conducive to victory for Black, not for Red." (Black makes a perfectly innocent move, which, thanks to a sequence of entirely unanticipated and wildly improbable coincidences, results in a threat to Red, which threat is then counteracted.) (iii) Standard cases of causation without counterfactual dependence—for example, cases of what Lewis calls "early cutting"—evoke not the slightest hesitation. Why should we think that failure of such dependence explains our hesitation with respect to the counterexamples?

So how might a proper defense of Lewis's counterintuitive "considered opinion" go? We have already seen how it ought to go, *in those cases which involve switching.* In fact, we saw that a defense much more detailed and sophisticated than Lewis's own is available, for

those cases. But here is the rub: the defense does not extend to cases that involve no switching. So the correct verdict about such cases is that they do have something to teach us about the metaphysics of causation, namely, that dependence does not suffice for causation. In turn, the route to that conclusion has something to teach us about the *methodology* of the metaphysics of causation: given that one has started the job of analyzing the structure of alleged counterexamples to some interesting philosophical thesis, one should take care to finish it. Lewis has not.

Have I finished it? Yes, as far as I can tell. That is, on careful reflection I cannot discern any further structure to the various examples which matters, or which distinguishes some of them from others. But, of course, such a conclusion is by its nature tentative. Supposing it correct, we can now measure the price of transitivity, at least in the current market. Put the point this way: we knew all along that counterfactual dependence is not transitive; that is what we learn in our first course on the semantics of this conditional. And that should have at least raised the suspicion that there was something overly optimistic in the view that counterfactual dependence bears a deep connection to causation in the way that the dependence thesis brings out. The price of transitivity is that we must finally face up to this suspicion, recognize that it was well founded, and therefore look elsewhere for the tools with which to analyze our central concept of causation.

NED HALL

Massachusetts Institute of Technology