Two Concepts of Causation

Ned Hall

§1 Introduction

Causation, understood as a relation between events, comes in at least two basic and fundamentally different varieties. One of these, which I call "dependence", is simply that: counterfactual dependence between wholly distinct events. In this sense, event **c** is a cause of (distinct) event **e** just in case **e** depends on **c**; that is, just in case, had **c** not occurred, **e** would not have occurred. The second variety is rather more difficult to characterize, but we evoke it when we say of an event **c** that it helps to *generate* or *bring about* or *produce* another event **e**, and for that reason I call it "production". Here I will articulate, defend, and begin to explore the consequences of this distinction between dependence and production. A synopsis:

After taking care of some preliminaries (§2), I will argue for the distinction in a slightly devious manner, by starting with a broad-strokes critique of counterfactual analyses of causation (§3). The reason for this approach is plain: Since I end up endorsing the simplest kind of counterfactual analysis—albeit only as an analysis of *one* kind of event-causation—it makes sense to pay some attention to the prospects for this and kindred analyses, and to examine why there is no hope of turning them into analyses of a *univocal* concept of event-causation. Specifically, my critique will aim to show that the best attempts to shore up counterfactual analyses in the face of well-known and stubborn counterexamples (involving certain kinds of overdetermination) rely on three general theses about causation:

• Transitivity: If event **c** is a cause of **d**, and **d** is a cause of **e**, then **c** is a cause of **e**.

• Locality: Causes are connected to their effects via spatiotemporally continuous sequences of causal intermediates.

• Intrinsicness: The causal structure of a process is determined by its intrinsic, non-causal character (together with the laws).

These theses—particularly the second and third—will require more discussion and elaboration, which will come in due time. For now, contrast them with the thesis that lies at the heart of all counterfactual analyses of causation:

• Dependence: Counterfactual dependence between wholly distinct events is sufficient for causation.

The simplest counterfactual analysis adds that dependence is *necessary* for causation. As a *general* analysis of causation, it fails for well-known reasons, which we will review shortly. Consequently, no counterfactual analyst that I am aware of endorses this necessary condition. But to my knowledge, all endorse the sufficient condition codified in the thesis of Dependence. Indeed, it is probably safe to say that Dependence is the cornerstone of every counterfactual analysis.

What is the trouble? Simply this: A hitherto ignored class of examples involving what I call "double-prevention" reveals deep and intractable tensions between the theses of Transitivity, Locality, and Intrinsicness, on the one hand, and Dependence, on the other (§4).

In §5, I'll add to my case by arguing that exactly parallel tensions divide the first three theses from the thesis of

• Omissions: Omissions—failures of events to occur—can both cause and be caused.

This thesis will also need further elaboration and discussion.

One immediate result is that counterfactual analyses are doomed to failure (unless, as I think, they are understood to be narrowly targeted at just one *kind* of event-causation): for they *need* the first three theses if they are to cope with the well-known counterexamples involving overdetermination, but they *cannot abide* these theses if they are to cope with the counterexamples involving double-prevention (or, for that matter, if they admit omissions as causes and effects).

Although important, this result is eclipsed by a more significant lesson that I will develop in §6. For the five theses I have mentioned are, I claim, all *true*. Given the deep and intractable tensions between them, that can only be because they characterize *distinct concepts of causation*. Events can stand in one kind of causal relation—dependence—for the explication of which the counterfactual analysis is perfectly suited (and for which omissions can be perfectly suitable relata). And they can stand in an entirely different kind of causal relation—production—which requires an entirely different kind of analysis (and for which omissions are *not* suitable relata). Dependence and Omissions are true of the first of these causal relations; Transitivity, Locality, and Intrinsicness are true of the second. I'll close §6 by defending this claim against some of the most obvious objections.

How are production and dependence to be analyzed? Dependence, I think, is easy; it is counterfactual dependence, nothing more nor less (with, perhaps, the proviso that counterfactual dependence itself can come in different varieties; see §7 for brief discussion). Production is trickier, and in §7 I'll offer a speculative proposal about its analysis, confined to the special case of deterministic laws that permit no

action at a temporal distance or backwards causation. But I'll say at once that I am much more confident of the propriety of the distinction than I am of this particular gloss on the "production" half of it.

I'll close, in §8, by suggesting some ways in which the distinction between production and dependence might be put to work, and by highlighting what I think are the most important bits of unfinished business.

§2 Preliminaries, and a brief methodological sermon

There are, in the literature, at least a dozen versions of a counterfactual analysis of causation that I am aware of. To attack them all, in detail, would require (to borrow an apt term from Tim Maudlin) a kind of philosophical "trench warfare" that only deeply committed partisans could find engaging. I'll confess to a taste for trench warfare, but I won't indulge it here. Instead, I will follow a different strategy, focusing my critique on the simplest counterfactual analysis, according to which causation is counterfactual dependence between wholly distinct events. It will be far more illuminating to explore the most basic problems for this analysis—along with the clearest and most plausible strategies for confronting them—than it would be to wind through the convolutions built into the multitude of more sophisticated variants.

In order to develop this critique as constructively as possible, we must avoid various methodological pitfalls. For that reason, it will be important to characterize, if only in a rough way, the causal relation which is the target of the counterfactual analysis. I take the analysis to concern the concept of causation as a transitive, egalitarian relation between localized, datable events. Let's look at the parts of this characterization in turn.

Begin with the relata. In understanding them to be *events*, I am taking sides on an issue that has seen much recent controversy.[1] I grant that there may be senses in which non-events—facts, properties, maybe even things—can cause and be caused; certainly we speak of event *types* as doing so, as when we say that lightning causes fires. All the same, I assume that there is a clear and central sense of "cause"—the one at issue here—in which causes and effects are always events. (In §6, I'll qualify this assumption slightly, suggesting that dependence, at least, can admit more kinds of relata.)

I will, furthermore, follow common practice by stretching ordinary usage of the term "event" to cover such things as, for example, the presence, at the appropriate time, of the oxygen and dry timber that combine with the lightning bolt to produce the forest fire. I will also take it for granted that we can adequately discern when two events fail to be *wholly distinct*—that is, when they stand in some sort of logical or mereological relationship that renders them unsuited to stand in *causal* relationships—and that we can tell when a description is too disjunctive or extrinsic to succeed in picking out an event. Without such assumptions, it is far too easy to make a hash of the simple analysis, and the analyses that build on it, by way of alleged counterexamples to the claim that counterfactual dependence is sufficient for causation (cf. Kim, 1973). Examples:

Suppose that I shut the door, and in fact slam it. We may have two events here—a shutting and a slamming—distinguished because the first could have happened without the second (if, for example, I had shut the door softly). But if I hadn't shut the door, I couldn't (and so wouldn't) have slammed it—so it seems that the analysis wrongly tells us that the shutting is a cause of the slamming.

Another case: Suzy expertly throws a rock at a glass bottle, shattering it. The shattering consists at least in part of many smaller and more localized events: first the glass fractures, then one shard goes flying off this way, another that way, and so on. If the shattering hadn't happened then none of its constituent events would have happened—so it seems that the analysis wrongly tells us that the shattering is a cause of them all.

A third case: Far away, Billy throws another rock, shattering a different glass bottle. Suppose there is a "disjunctive" event **c** which, necessarily, occurs iff either Suzy's throw or Billy's throw occurs. If **c** hadn't occurred, neither bottle would have shattered—so it seems that, according to the analysis, we have discovered a fairly immediate common cause of these widely separated events.

A final case: Suppose there is a type of extrinsically specified event that, necessarily, has an instance occurring at time t iff I sent an email message exactly 2 days before t. Let **c** be such an instance, and let **e** be the reply, occurring at, say, a time one day after I sent the given email (and so a day before **c**). Since **e** would not have happened if **c** hadn't, it seems that, according to the analysis, we have backwards causation very much on the cheap.

Each case deserves a different diagnosis. In the first case, we should say that the events fail to be distinct because of their logical relationship. In the second case, we should say that the shattering is not distinct from its constituents, because of their mereological relationship. In the third case, we should say that the disjunctive event is not an event at all, hence not apt to cause (or be caused). In the final case, we

should say that there are no such extrinsically specified event types. We should say all these things, and it's up to a philosophical theory of events to tell us why we are justified in doing so.[2]

Of course, I do not at all mean to suggest that it is an easy matter to provide an adequate philosophical account of events that meets these criteria. I certainly won't try to provide any such account here. What I *will* do is avoid choosing examples where any of the controversies surrounding the nature of events makes a difference.[3]

Turn next to the characterization of the relation. Transitivity is straightforward enough: if event **a** is a cause of event **b**, and **b** a cause of **c**, then **a** is thereby a cause of **c**. What I mean by "egalitarian" can best be made clear by contrast with our usual practice. When delineating the causes of some given event, we typically make what are, from the present perspective, invidious distinctions, ignoring perfectly good causes because they are not sufficiently salient. We say that the *lightning bolt* caused the forest fire, failing to mention the contribution of the oxygen in the air, or the presence of a sufficient quantity of flammable material. But in the egalitarian sense of "cause", a complete inventory of the fire's causes must include the presence of oxygen and of dry wood. (Note that transitivity helps make for an egalitarian relation: events causally remote from a given event will typically not be *salient*—but will still be among its causes, for all that.)

Now for a brief methodological sermon: If you want to make trouble for an analysis of causation—but want to do so on the cheap—then it's convenient to ignore the egalitarian character of the *analysandum*. Get your audience to do the same, and you can proceed to elicit judgments that will appear to undermine the analysis, but which are in fact irrelevant to it. Suppose that my favorite analysis counts the Big Bang as among the causes of today's snowfall (a likely result, given transitivity). How easy it is to refute me, by observing that if asked what *caused* the snowfall (better still: what was *the* cause of it), we would never cite the Big Bang! Of course, the right response to this "refutation" is obvious: It conflates the transitive, egalitarian sense of "cause" with a much more restrictive sense (no doubt greatly infected with pragmatics) that places heavy weight on salience.

A simple mistake, it would seem. But the same sort of mistake shows up, in more subtle forms, in examples drawn from the literature. It will be helpful to work through a few illustrative cases—ones that show, incidentally, how even first-rate authors can sometimes go astray.

First, Bennett (1987, pp. 222-3; italics in the original), who is here concerned with Lombard's thesis that an event's time is essential to it:

> Take a case where this is true:
>
> > There was heavy rain in April and electrical storms in the following two months; and in June the lightning took hold and started a forest fire. *If it hadn't been for the heavy rain in April, the forest would have caught fire in May.*
>
> Add Lombard's thesis to that, and you get
>
> > If the April rain hadn't occurred the forest fire wouldn't have occurred.
>
> Interpret that in terms of the counterfactual analysis and you get
>
> > The April rains caused the forest fire.
>
> That is unacceptable. A good enough theory of events and of causation might give us reason to accept some things that seem intuitively to be false, but no theory should persuade us that delaying a forest's burning for a month (or indeed for a minute) is causing a forest fire.

Lombard agrees that Bennett's result "is unacceptable. It is a bit of good common sense that heavy rains can put out fires, they don't start them; it *is* false to say that the rains caused the fire." (Lombard 1990, p. 197; italics in the original)

Lombard discusses a second example which shows that the essentiality of an event's time is not at issue (*ibid.*, pp. 197-8):

> Suppose that Jones lives in a very dangerous neighborhood, and that one evening Smith attempts to stab him to death. Jones is saved because of the action of Brown who frightens Smith off. However, a year later, Jones is shot to death by the persistent Smith. So, if Brown's action had not occurred, Jones's death due to the shooting would not have occurred, since he would have died of stab wounds a year earlier. But, I find it intuitively quite unacceptable to suppose that Brown's action was a cause of Jones's dying as a result of gunshot a year later.

Finally, Lewis discusses a very similar example (Lewis 1986b, p. 250):

> It is one thing to postpone an event, another to cancel it. A cause without which it would have occurred later, or sooner, is not a cause without which it would not have occurred at all. Who would dare be a doctor, if the hypothesis under consideration [that an event's time is essential to it] were right? You might manage to keep your patient alive until 4:12, when otherwise he would have died at 4:08. You would then have caused his death. For his death was, in fact, his death at 4:12. If that time is essential, his death is an event that would not have occurred had he died at 4:08, as he would have done without your action. That will not do.

If these examples are meant to provide rock-solid "data" upon which the counterfactual analysis (and perhaps others) founders, then they uniformly fail—for in each case, we can find *independently plausible* premises that entail the allegedly unacceptable consequences. Of course that doesn't show that the consequences are *true.* But it *does* show that we make a serious methodological mistake if we treat those of our intuitions that run counter to them as non-negotiable "data".

First we must disentangle irrelevant but confusing issues. It is probably right that an event's time is not in every case essential to it; but (*pace* Lewis) that doesn't help in any of the three cases. This is more or less obvious in the first two cases (the June fire is not the same as the fire that would have happened in May; the death by shooting is not the same as the death by stabbing that would have happened a year earlier). So consider Lewis's case. Supposedly, it "will not do" to assert that the doctor's action is among the causes of the patient's death. But what does this have to do with the proximity of the actual time of death to the time at which the patient would have died? Suppose you manage to keep your patient alive until June of 1999, when otherwise he would have died in June of 1998. Would you then have caused his death, since without your action the death he in fact died would not have occurred? It is no less (and no more) unacceptable to say "yes" in this case than it is to say "yes" in Lewis's case. But if, following Lewis, we conclude that the actual death is the same as the death which would have occurred a year earlier, then we are taking the denial of the essentiality of times to a ridiculous extreme. Such a denial, however warranted, does not give the counterfactual analyst the means to respond effectively even to Lewis's problem.

The analyst can, however, draw on our brief methodological sermon to point to two sorts of judgments about causation which the three examples implicitly trade on—but illegitimately, since these judgments concern types of causation which are not at issue. We can all agree that "heavy rains can put out fires, they don't start them," just as we can agree that smoking causes lung cancer, but regular exercise doesn't. So what? The intuitions called upon here do not concern the concept of causation as a transitive, egalitarian relation between events, but rather some other concept of causation as an inegalitarian relation between event-*types*. (Never mind that *starting* a fire is not the only way to be one of its causes!) Similarly, we can all agree that it is the *lightning* that causes the forest fire, and nothing else—including the heavy rains. Again, so what? Here we seem to have in mind a restricted, inegalitarian concept of event-causation according to which events that are to count as causes must be particularly salient in some respect; but judgments involving *this* concept matter not at all to the counterfactual analysis, since it concerns the weaker and more inclusive transitive, egalitarian concept.

Unfortunately, Bennett, Lombard and Lewis have all muddied the question of whether the counterfactual analysis is adequate by choosing examples where intuitions of the two types just discussed are particularly strong and *seemingly* salient: It's not the *rainfall* that causes the June fire, but rather the lightning; moreover, it's just "good common sense" that heavy rains don't cause fires![4] Of course, while recognizing these points you might still judge these cases to have some intuitive force as counterexamples. Fair enough; they do. But a more careful examination shows how hasty it would be to take any such intuitions as decisive. I'll make the case against Bennett, after which it will be clear enough how to proceed against Lombard and Lewis that we can leave those cases aside.

The idea is to find an event intermediate between the cause and its alleged effect which is *clearly* a cause of the second, and at least plausibly an effect of the first. So observe that among the causes of the June fire is not just the lightning but also the very presence of the forest, filled with flammable material. The presence of the forest in the hours before the lightning strikes is itself an event, or perhaps a collection of events. This event is a cause of the June fire (albeit not a salient cause). What are *its* causes? A typical counterfactual analysis will claim that one of its causes is the April rainfall, since without the rainfall the forest would have been destroyed in May. But we can argue for the plausibility of this claim independently, by noting that the following judgments seem, intuitively, to be correct: it is in part *because of* the April rains that the forest is present in June; any complete *causal explanation* of the forest's presence must cite the role of the April rains in preventing its destruction; the April rains are at least in part *responsible for* the presence of the forest in June.[5]

One could deny the truth of these judgments, or deny that they show that the April rainfall is a cause of the forest's presence in June, or deny that causation is transitive in the way that is needed to complete the inference to the claim that the April rainfall is among the causes of the June forest fire. But unless one can find some grounds for supporting such denials—grounds independent of the mere intuitive implausibility of the claim in question—then this implausibility will fail to provide a particularly compelling reason for giving up the counterfactual analysis. (Exactly parallel points apply to the other two examples.) Happily, I think the conclusions drawn in §6 clear up what is going on in the rainfall case (and the other cases), precisely by showing that counterfactual analyses utterly fail to capture one important sense of "cause"—production—and that in this sense the April rains are *not* among the causes of the June fire. But it will take some work to get there, and along the way we must not be distracted by the temptations of such bogus "refutations" as those we have just examined. Intuitions about cases must be heeded, to be sure. But not blindly.

Onward. It will help to have a means of representing simple causal structures; accordingly, I will adopt the "neuron" diagrams used by Lewis.[6]

The diagram below depicts a pattern of neuron firings. Gray circles represent firing neurons, while arrows represent stimulatory connections between neurons. The order of events is left to right: In figure 1 neuron **a** fires, sending a stimulatory signal to neuron **b**, causing it to fire; **b**'s firing in turn sends a stimulatory signal to neuron **c**, causing it to fire.

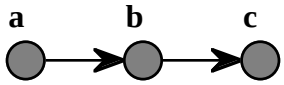
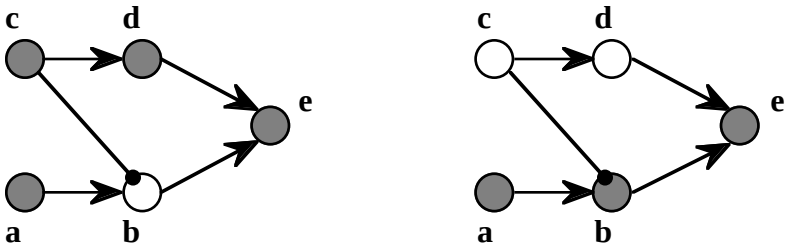


Figure 1

We will also need a way to represent *prevention* of one event by another. So let us add inhibitory connections to the neuron diagrams, represented by a line ending in a solid dot:



c fires if **c** hadn't fired…

Figure 2

In the left-hand diagram, neurons **a** and **c** fire simultaneously; **c**'s firing causes **d** to fire, which in turn causes **e** to fire. However, thanks to the inhibitory signal from **c**, **a**'s firing does *not* cause **b** to fire; **b**'s failure to fire is represented by leaving its circle unshaded. The right-hand diagram shows what would have happened if **c** had not fired.

Calling these diagrams "neuron diagrams" is merely picturesque; what is important about them is that they can provide, in a readily digestible form, partial representations of many causal structures.

§3 The simple counterfactual analysis, and two kinds of overdetermination

§3.1 The simple analysis

Both for simplicity and to avoid needless controversies, I will focus only on the counterfactual analysis as it applies to worlds with *deterministic* laws that permit neither backwards causation nor action at a temporal distance (although the lessons of the paper apply much more generally, as far as I can see). I will also leave aside cases, if such there be, where a cause is simultaneous with one of its effects.

Here and throughout I will denote events by lower-case bold-face letters '**a**', '**b**', '**c**', and so forth; the proposition that an event **e** occurs by 'O**e**'; and the counterfactual or subjunctive conditional by '$\rightarrow$' (read: "were it the case that … then it would be the case that …").[7] The simple analysis is as follows:

Event **c** is a cause of event **e** iff

(i) **c** and **e** are wholly distinct events;

(ii) O**c**, O**e**, and ~O**c** → ~O**e** are all true.

An immediate problem arises, whose solution requires the counterfactual conditional to be understood in a rather specific way. In figure 1, it is certainly correct to say that if **a** hadn't fired, **c** wouldn't have; but it may also be correct to say that if **c** hadn't fired, **a** wouldn't have. (If you don't like the sound of that, try the happier locution: If **c** hadn't fired, it would have to have been that **a** didn't fire.) If so, the analysis wrongly says that **c** is a cause of **a**. (Note the harmless but convenient ambiguity: letters sometimes refer to events, sometimes to neurons.)

Two responses seem possible. We could augment the analysis by adding some third condition to guarantee the asymmetry of causation (for example: causes must precede their effects). Or we could deny the truth of the offending counterfactual, appealing to an account of the conditional which secured the falsehood of all such "backtrackers" (to use Lewis's apt term).[8] Swain (1978), for example, opts for the first alternative, Lewis (1979 and 1986a) (and most other counterfactual analysts) for the second.[9]

The first response doesn't work, partly for reasons that have been well-explored and that I won't rehearse in detail here (e.g., merely adding the requirement that causes precede their effects won't help if, say, **c** and **e** are joint effects of some event **a**, with **c** occurring before **e**; for we could still reason that if **c** hadn't happened, it would have to have been that **a** didn't happen, and therefore that **e** didn't happen). A different, often unnoticed reason for rejecting the first response deserves some discussion, however: The problem is that this response implicitly supposes that backtrackers threaten only the *sufficiency* of the above analysis. If that were true, it would make sense to add further conditions, so as to make the analysis less liberal. But backtrackers also undermine its *necessity*, as figure 2 shows.

In figure 2, **d** is, clearly, a cause of **e**. But if, in evaluating counterfactuals with the antecedent "**d** does not occur", we proceed by making minimal alterations to the past events that led to **d**, then we will reach a counterfactual situation in which **c** does not occur, but **a** still does—that is, a counterfactual situation in which **e** occurs. That is, if we allow as true the backtracker ~O**d** → ~O**c**, then the right-hand diagram *also* describes what would have happened if **d** hadn't fired, and so the conditional ~O**d** → ~O**e** is false. Then how can it be that **d** turns out to be a cause of **e**? Adding *extra* conditions to (i) and (ii) provides no answer.[10] (Nor will it help to liberalize the analysis in the standard way, by taking causation to be the *ancestral* of counterfactual dependence. For the problem that threatens the connection between **e** and **d** will equally threaten the connection between **e** and any event that mediates between **d** and **e**.)

In short, reading the counterfactual in a backtracking manner destroys the dependence of **e** on **d**. That's not only trouble for the simple analysis: it's just wrong, since it manifestly *is* the case that if **d** hadn't fired, **e** wouldn't have. Or, more cautiously, there manifestly *is* an acceptable reading of the counterfactual conditional according to which this is true. I will henceforth take it for granted that both the simple analysis and its more elaborate kin employ such a "non-backtracking" reading of the conditional.

Note a crucial feature of this reading, however. Specifically, we don't avoid the problem raised by figure 2 merely by denying the backtracking conditional ~O**d** → ~O**c**. For to deny a counterfactual X → Y is *not* to assert the contrary conditional X → ~Y, but rather to assert the weaker "might" conditional, symbolized as X ◊→ ~Y and read "had X been true, Y *might have been* false."[11] This "might" conditional is weaker because, unlike X → ~Y, it is *consistent* with the "might" conditional X ◊→ Y. In the case at hand, the falsehood of ~O**d** → ~O**c** is consistent with the truth of ~O**d** ◊→ ~O**c**—in other words, consistent with the claim that if **d** had not fired, then a course of events which *might have occurred* is that described by the right-hand diagram. But in that case, if **d** had not fired, then **e** might have fired anyway; in symbols, ~O**d** ◊→ O**e**. And this "might" conditional is the denial of the conditional ~O**d** → ~O**e**. So, not only must the conditional ~O**d** → ~O**c** be false, but the conditional ~O**d** → O**c** must be true. If **d** hadn't fired, **c** would have fired just the same.

It is a matter of some controversy what is the proper semantics for this kind of conditional. Recall the standard possible worlds semantics: The conditional X → Y is true iff some possible world where X and Y are both true is "closer" to actuality than any world where X is true but Y false. Without trying to define the "closer than" relation, we can still establish this much about it: Some world w where **c** fires but **d** does not is closer to actuality (the left-hand diagram) than any world where neither **c** nor **d** fires (for if **d** had not fired, **c** still would have). How can this be? Don't the laws guarantee that if **c** fires, then **d** fires?

Various responses are possible. One might point out that **c** alone is not lawfully sufficient for **d**: other conditions must conspire with the firing of **c** to bring about the firing of **d**, and in w (so this story goes) some of those other conditions aren't met. Or one might deny that the laws of our world hold without exception in w. Of course, they had better hold *almost* without exception—in particular, they had better hold from the time of **d**'s (non)occurrence forward (else the evaluation of the conditional ~O**d** → ~O**e** goes haywire). But that is no problem: we need only admit just enough of a violation to break the

connection between **c** and **d**.[12] Or one might deny that a *uniform* semantics is required for conditionals X $\rightarrow$ Y that are "forward tracking" (in the sense that X concerns times before those that Y concerns) and those that are backtracking.[13] Fortunately, we need only come up with a rule for evaluating counterfactuals of the form ~O**c** $\rightarrow$ ~O**e**, where **c** and **e** both occur, and **c** precedes **e**—and on such a rule, the three foregoing approaches can certainly agree. Following Maudlin (2000a), I suggest the following: Letting t be the time of occurrence of the given event **c**, we evaluate the conditional ~O**c** $\rightarrow$ ~O**e** by altering the state of the world at t just enough to make the antecedent true (without regard to what the past would have to have been like in order to give rise to that counterfactual t-state), evolving that state forward in time in accordance with the (actual) laws, and seeing whether the consequent comes out true.[14] So, in figure 2, if **d** hadn't fired, circumstances contemporaneous with its firing such as the non-firing of **b** would have been unchanged, and so **e** would not have fired. We can leave it undecided what the past would have been like, or even whether the semantics for our conditional needs to answer that question.

§3.2Early pre-emption

The simple analysis may appear quite able to stave off challenges to its *sufficiency*. But obvious problems beset its claim to necessity. Consider a case of ordinary pre-emption, as in figure 2. The firing of **e** is overdetermined by the simultaneous firings of **a** and **c**. But not in a way that leaves us at all uncertain as to what causes what: Without question, **c** is a cause of **e**, even though if **c** hadn't occurred, **e** would have occurred anyway, thanks to an alternative process, beginning with **a**, which **c** pre-empts.

There is an obvious strategy for handling this kind of case. First, we liberalize our analysis, by taking causation to be the *ancestral* of counterfactual dependence: **c** is a cause of **e** iff there are events $\mathbf{d}_1$, …, $\mathbf{d}_n$ such that $\mathbf{d}_1$ counterfactually depends on **c**, $\mathbf{d}_2$ depends on $\mathbf{d}_1$, …, and **e** depends on $\mathbf{d}_n$. Next, we look for an event **d** (or sequence of events) intermediate between the preventing cause **c** and the effect **e**, such that **e** depends on **d** and likewise **d** on **c**. The strategy works handily in the case before us (provided, once again, that we are careful *not* to interpret the counterfactual in a "backtracking" sense, according to which, had **d** not fired, it would have to have been the case that **c** didn't fire, and so **b** would have fired, and so **e** would still have fired).

Observe how natural this embellishment to the simple analysis is—and observe that it gives a central role to the Transitivity thesis.

§3.3Late pre-emption

Other, quite ordinary cases of overdetermination require a different treatment. Consider a case of so-called "late pre-emption", as in figure 3:

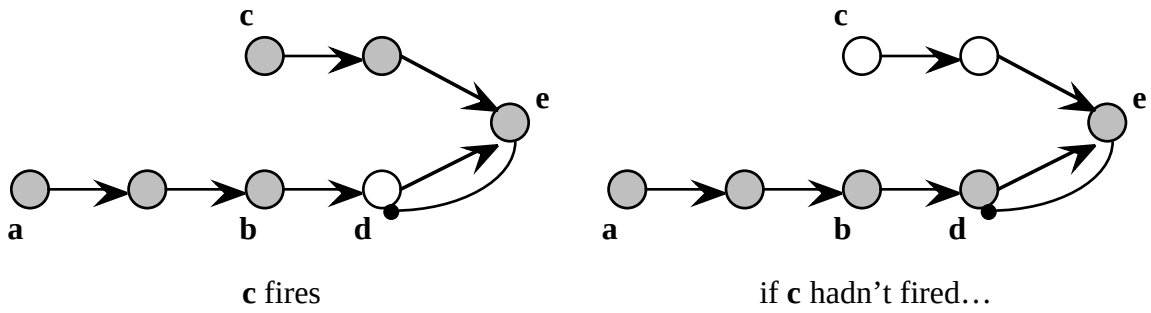


Figure 3
(neurons **a** and **c** fire simultaneously)

Neurons **a** and **c** fire simultaneously, so that **e** fires at the same time as **b**; the inhibitory signal from **e** therefore prevents **d** from firing. If **c** hadn't fired, **e** still would have; for in that case **d** would not have been prevented from firing and so would have stimulated **e** to fire. Likewise for *every* event in the causal chain leading from **c** to **e**: if that event had not occurred, **e** would nevertheless have fired. So the strategy of finding suitable intermediates breaks down; for it to succeed, **e** would have to depend on at least *one* event in the chain leading back to **c**, and it does not.

Here is another example with a slightly different structure; it illustrates how absolutely mundane these cases are. Suzy and Billy, our expert rock-throwers, are engaged in a competition to see who can shatter a target bottle first. They both pick up rocks and throw them at the bottle, but Suzy throws hers a split second before Billy. Consequently Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's had not occurred, so the

shattering is overdetermined. Once the bottle has shattered, however, it cannot do so again; thus the shattering of the bottle prevents the process initiated by Billy's throw from itself resulting in a shattering.

Suzy's throw is a cause of the shattering, but Billy's is not. Indeed, every one of the events which constitute the trajectory of Suzy's rock on its way to the bottle is a cause of the shattering. But the shattering depends on none of these events, since had any of them not occurred the bottle would have shattered anyway, thanks to Billy's expert throw. So the transitivity strategy fails.

Three alternative strategies for dealing with this kind of case suggest themselves. The first rests on the observation that Suzy's throw makes a difference to the time and manner of the shattering, whereas Billy's does not. The second rests on the observation that Suzy's throw is connected to the shattering by a spatiotemporally continuous chain of causal intermediates, whereas Billy's is not. And the third rests on the observation that there is a sequence of events connecting Suzy's throw to the shattering that has the right sort of *intrinsic* character to count as a causal sequence, whereas no such sequence connects Billy's throw to the shattering. Let us consider these strategies in turn.

There are various ways to implement the first strategy. For example, we could deny that the effect which does the preventing is numerically the same as the effect which would have occurred via the alternative process; if so, then our two examples *do* exhibit the needed pattern of counterfactual dependence, since the effect which actually occurred would not have occurred without its cause (although a very similar event would have occurred in its place). In figure 3, the firing of **e** which *would have* occurred, had **c** not fired, is not the same event as the firing which *actually* occurs. Likewise, if Suzy had not thrown her rock then the bottle would still have shattered—but later than it actually did, so it would not have been the same shattering. Alternatively, we could remain silent about the individuation of events, and simply employ a slightly different counterfactual in the analysis—say, by counting **c** a cause of **e** iff, had **c** not occurred, **e** would not have occurred at the time it *actually* did (Paul 1998). In recent work, Lewis (2000) has argued that we should count **c** a cause of **e** if there is a suitable pattern of counterfactual dependence between various different ways **c** or something like it might have occurred and correspondingly different ways in which **e** or something like it might have occurred. (Lewis proposes taking causation itself to be the ancestral of this relation.)

These approaches are uniformly non-starters. Never mind the well-known problems (e.g., that non-causes can easily make a difference to the time and manner of an event's occurrence—a gust of wind that alters the course of Suzy's rock ever so slightly, for example). What seems to have gone unnoticed is that it is not at all essential to examples of late pre-emption that the genuine cause make *any* difference to the time or manner of the effect. As Steve Yablo pointed out to me, it's easy enough to construct cases in which **c** is clearly a cause of **e**, but in which neither **c** nor any event causally intermediate between it and **e** makes the slightest difference to the way **e** occurs. Yablo points out that we can simply alter the story of Billy and Suzy. This time, Billy throws a Smart Rock, equipped with an on-board computer, exquisitely designed sensors, a lightning-fast propulsion system—and instructions to make sure that the bottle shatters in exactly the way it does, at exactly the time it does. In fact, the Smart Rock doesn't need to intervene, since Suzy's throw is just right. But had it been any different—indeed, had her rock's trajectory differed in the slightest, at any point—the Smart Rock would have swooped in to make sure the job was done properly. Sure, the example is bizarre. But not in a way that matters in the slightest to the evaluation of the causal status of Suzy's throw: Smart Rock notwithstanding, her throw is *still* a cause of the shattering—even though neither it nor any event that mediates between it and the shattering makes a difference to the time or manner of that shattering.

I won't consider these approaches further. It will be far more instructive for us to focus on the two alternative strategies.

Suzy's throw is spatiotemporally connected to the shattering in the right way, but Billy's is not. So perhaps we should add the Locality thesis as a constraint on the analysis: Causes have to be connected to their effects via spatiotemporally continuous sequences of causal intermediates. Now, on the face of it this is a step in entirely the *wrong* direction, since it makes the *analysans* more stringent. But if we simultaneously liberalize the analysis in other respects, this strategy might work. For example, we might say that **c** is a cause of **e** just in case there is a spatiotemporally continuous sequence of events connecting **c** with **e** and a (possibly empty) set S of events contemporaneous with **c** such that each later event in the sequence (including **e**) depends on each earlier event—or at least *would have*, had the events in S not occurred. That will distinguish Suzy's throw as a cause, and Billy's as a non-cause.

Of course, since action at a distance is surely *possible*, and so Locality at best a highly interesting contingent truth, this amended counterfactual analysis lacks generality. But it is patently general enough to be of value. At any rate, it is not so important for our purposes whether this strategy, or some variant,

can handle all cases of late pre-emption. What is important is that it is a plausible and natural strategy to pursue—and it gives a central role to the Locality thesis.

Lewis has proposed a third, different strategy. He begins with the intuition that the causal structure of a process is intrinsic to it (given the laws). As he puts it:

> Suppose we have processes—courses of events, which may or may not be causally connected—going on in two distinct spatiotemporal regions, regions of the same or of different possible worlds. Disregarding the surroundings of the two regions, and disregarding any irrelevant events that may be occurring in either region without being part of the process in question, what goes on in the two regions is exactly alike. Suppose further that the laws of nature that govern the two regions are exactly the same. Then can it be that we have a causal process in one of the regions but not the other? It seems not. Intuitively, whether the process going on in a region is causal depends only on the intrinsic character of the process itself, and on the relevant laws. The surroundings, and even other events in the region, are irrelevant.

In cases of late pre-emption, the process connecting cause to effect does not exhibit the right pattern of dependence—but only because of accidental features of its surroundings. The process that begins with Suzy's throw and ends with a shattered bottle does not exhibit the right pattern of dependence (thanks to Billy's throw), but it is intrinsically just like other possible processes that do (namely, processes taking place in surroundings that lack Billy, or a counterpart of him). Lewis suggests, in effect, that *for that reason* Suzy's throw should count as a cause.

Clearly, Lewis is trying to parlay something like the Intrinsicness thesis into an amended counterfactual analysis, one adequate to handle cases of late pre-emption. Now, I think there are serious problems with the details of Lewis's own approach (spelled out in the passage following that just quoted), but since that way lies trench warfare, I won't go into them. I do, however, want to take issue with his statement of the Intrinsicness thesis, which is too vague to be of real use. What, after all, is a "process" or "course of events"? If it is just any old sequence of events, then what he says is obviously false: We might have a sequence consisting of the lighting of a fuse, and an explosion—but whether the one is a cause of the other is not determined by the intrinsic character of this two-event "process", since it obviously matters whether *this* fuse was connected to *that* exploding bomb.

I will simply give what I think is the right statement of the Intrinsicness thesis, one which eschews undefined talk of "processes".[15] Suppose an event **e** occurs at some time t'. Then consider the structure of events which consists of **e**, together with all of its causes back to some arbitrary earlier time t. That structure has a certain intrinsic character, determined by the way the constituent events happen, together with their spatiotemporal relations to one another. It also has a certain causal character: in particular, each of the constituent events is a cause of **e** (except **e** itself, of course). Then the Intrinsicness thesis states that any possible structure of events that exists in a world with the same laws, and that has the same intrinsic character as our given structure, *also* duplicates this aspect of its causal character—that is, each duplicate of one of **e**'s causes is itself a cause of the **e**-duplicate.[16]

Three observations: First, "same intrinsic character" can be read in a very strict sense, according to which the two structures of events must be *perfect* duplicates. Read this way, I think the Intrinsicness thesis is incontrovertible. But it can also be read in a less strict sense, according to which the two structures must be, in some sense, sufficiently *similar* in their intrinsic characters. Read this way, the thesis is stronger but still highly plausible. Consider again the case of Billy and Suzy, and compare the situation in which Billy throws his rock with the situation in which he doesn't. Clearly, there is a strong intuition that the causal features of the sequence of events beginning with Suzy's throw and ending in the shattering should be the *same* in each case, precisely *because* Billy's throw is extrinsic to this sequence. But it is too much to hope for that the corresponding sequences, in each situation, be *perfect* duplicates; after all, the gravitational effects of Billy's rock, in the situation where he throws, will make minute differences to the exact trajectory of Suzy's rock, etc. So if it is the Intrinsicness thesis that gives voice to our conviction that, from the standpoint of Suzy's throw, the two situations must be treated alike, then we should read the "same intrinsic character" clause in that thesis in the less stringent way.

Doing so quite obviously leaves us with the burden of explaining what near-but-not-quite-perfect duplication of intrinsic character consists in. I won't try to unload that burden here. It will emerge that for my *main* purposes, that doesn't matter, since in order to use the Intrinsicness thesis to argue that dependence and production are two distinct kinds of causation, I can read "same intrinsic character" in the more stringent sense. (Alas, we will also see that my own preferred *analysis* of production will require the less stringent reading.)

The second observation to make about the Intrinsicness thesis is that it is somewhat limited in scope: it does not apply, in general, to situations in which there is causation at a temporal distance, or to situations in which there is backwards causation. Roughly, the problem is that the relevant structure of events must be *complete* in a certain respect, consisting in a complete set of joint causes of the given effect **e**, together with all of those events that mediate between these causes and **e**. I won't go into the reasons why it must exhibit this kind of completeness (but see my 2000c). But consider a case where the effect takes place at 1 o'clock, and we have collected together all of its causes that occur at noon, as well as those that occur between noon and 1. If there is action at a temporal distance, then some of the other causes with which the noon causes combine to bring about the effect might have occurred *before* noon, in which case our structure won't be sufficiently complete. If there is backwards causation, then some of the events that mediate between the noon causes and the effect might occur *outside* the given interval, in which case our structure won't be sufficiently complete. Either way, there is trouble. It is partly in order to finesse this trouble that I have limited my focus by ignoring both backwards causation and causation at a temporal distance.

The third observation to make about the Intrinsicness thesis is that we must assume—on pain of rendering the thesis trivially false—that the structure of events against which we compare a given structure includes no *omissions*. Let the structure S consist of **e**, together with all of its causes back to some arbitrary earlier time t.. And let the structure S' simply consist of S, together with some arbitrary omission that "occurs" at some point in the relevant interval. Plausibly, this omission will contribute nothing to the intrinsic character of S'—for it simply consists in the *failure* of some type of genuine event to occur. So S' will perfectly match S. If we apply the Intrinsicness thesis uncritically, we immediately get the absurd result that the added omission—whatever it is!—counts as a cause of **e**. Now, it was already fairly clear that whatever the guiding intuition is behind the Intrinsicness thesis, it does not concern omissions. This result confirms the suspicion. So the final clause of the Intrinsicness thesis should read: "…any possible structure of *genuine* events (not including any omissions) that exists in a world with the same laws, and that has the same intrinsic character as our given structure, *also* duplicates…." (It doesn't follow that S—the structure picked out as consisting of **e**, together with all of its causes back to some earlier time t—must include no omissions. We'll take up the question of whether it can in §5, below.)

Perhaps the counterfactual analyst can use the Intrinsicness thesis to handle the problem of Billy and Suzy. After all, in the alternative circumstances in which Billy's throw is absent, it seems correct to say that the causal history of the shattering (back to the time of Suzy's throw) consists exactly of those events on which it depends. What's more, this structure matches a structure that takes place in the actual circumstances, where Billy's throw confounds the counterfactual relations; Suzy's throw, being a part of this structure, will therefore count as a cause of the shattering, thanks to the Intrinsicness thesis. To be sure, this is no more than a suggestion of a revised analysis. But again, what is important is that it is a plausible and natural suggestion to pursue—and it gives a central role to the Intrinsicness thesis.

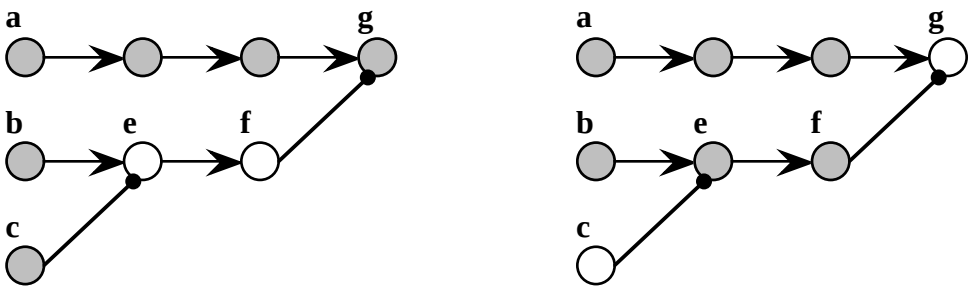
§4 Double-prevention

Now for something completely different: a kind of example that spells trouble for the *sufficiency* of the simple analysis, by showing that the cornerstone thesis of Dependence runs headlong into conflict with each of Transitivity, Locality, and Intrinsicness.

§4.1 Example

Suzy and Billy have grown up, just in time to get involved in World War III. Suzy is piloting a bomber on a mission to blow up an enemy target, and Billy is piloting a fighter as her lone escort. Along comes an enemy fighter plane, piloted by Enemy. Sharp-eyed Billy spots Enemy, zooms in, pulls the trigger, and Enemy's plane goes down in flames. Suzy's mission is undisturbed, and the bombing takes place as planned. If Billy hadn't pulled the trigger, Enemy would have eluded him and shot down Suzy, and the bombing would not have happened.

This is a case of what I call "double prevention": one event (Billy's pulling the trigger) prevents another (Enemy's shooting down Suzy), which had it occurred would have prevented yet another (the bombing). The salient causal structure is depicted in figure 4:



c fires if **c** hadn't fired…

Figure 4

Neurons **a**, **b**, and **c** all fire simultaneously. **c**'s firing prevents **e** from firing; if **e** had fired, it would have caused **f** to fire, which in turn would have prevented **g** from firing. Thus, if **c** had not fired, **g** would not have. So **c** is a cause of **g**: Billy's pulling the trigger is a cause of the bombing.

This consequence of the counterfactual analysis might seem natural enough. After all, wouldn't we give Billy part of the credit for the success of the mission? Isn't Billy's action part of the explanation for that success? And so on. On the other hand, it might seem quite unnatural—for the scuffle between Billy and Enemy takes place, let us suppose, hundreds of miles away from Suzy, in such a way that not only is she completely oblivious to it, but it has absolutely no effect on her whatsoever. Here she is, in one region, flying her plane on the way to her bombing mission. Here Billy and Enemy are, in an entirely separate region, acting out their fateful drama. Intuitively, it seems entirely unexceptionable to claim that the events in the second region have no impact at all on the events in the first—for isn't it plain that no physical connection unites them?

So far, it might seem that we have a stalemate: two contrary intuitions about the case, with no way to decide between them. (Indeed, my informal polling suggests that intuitive judgments vary quite a lot.) Not so: Both the judgment that we have a case of causation here, and the thesis of Dependence which endorses this judgment, run into trouble with each of the theses of Locality, Intrinsicness, and Transitivity.

§4.2Problems with Locality

We all know what action at a distance is: we have a case of it if we have a cause, at least one of whose effects is not connected to it via any spatiotemporally continuous causal chain.[17] I take it that action at a distance is possible, but that its manifestation in a world is nevertheless a highly non-trivial fact about that world. Yet if Billy's action counts as a cause of the bombing, then the quite ordinary and mundane relationship it bears to the bombing also counts as a case of action at a distance. Is *this* all it takes to achieve non-locality? (And to think that philosophers have been fussing over Bell's Inequalities!) If so, we would be hard-pressed to describe laws that *didn't* permit action at a distance. For example, even the classical laws that describe perfectly elastic collisions would have to be judged "non-local", since they permit situations in which one collision prevents a second, which, had it happened, would have prevented a third—so that we have dependence of the third collision on the first, but no connecting sequence of causal intermediates. In short, it appears that while Dependence doesn't *quite* contradict Locality, it renders it satisfiable only by the most trivial laws (e.g., laws which say that nothing ever changes). That's wrong: The distinction between laws which do and laws which don't permit action at a distance is interesting; to assimilate it to the all-but-vacuous distinction between laws which do and laws which don't permit double-prevention is a mistake.

A remarkably frequent but entirely unsatisfactory response is the following: Billy's action *is* connected to the bombing via a spatiotemporally continuous causal chain—it's just that this chain consists, in part, of *omissions* (namely, the various failures of Enemy to do what he would have done, had Billy not fired). Now, it's not just that such reliance on causation by omission is desperate on its face (though indeed it is). It's that even if we grant that these omissions exist, and are located where the events omitted would have occurred (a non-trivial supposition: right now I am at home, and hence fail to be in my office; is this omission located there or here?), it doesn't help. For there is no reason to believe that the region of spacetime these omissions occupy intersects the region of spacetime that Suzy and her bomber *actually* occupy; to hold otherwise is just to mistake *this* region with the region she *would have* occupied, had Billy not fired. We can agree that had Billy not fired, then the Enemy-region would have intersected the Suzy-region; but if, say, Suzy would have swerved under those circumstances, then it's

just false to suppose that this counterfactual Enemy-region (= the *actual* omission-of-Enemy-region) intersects the *actual* Suzy-region.

Of course, the debate can take various twists and turns from here: there are further stratagems one might resort to in an effort to interpolate a sequence of omissions between Billy and the bombing; alternatively, one might deny that causation without a connecting sequence of causal intermediates really *is* sufficient for action at a distance. It won't profit us to pursue these twists and turns (but see my 2000b); suffice it to say that the stratagems fail, and the prospects for a replacement for the sufficient condition seem hopeless.

§4.3Problems with Intrinsicness

Let's first recall what the Intrinsicness thesis says, in its careful formulation: Suppose an event **e** occurs at some time t'. Consider the structure of events S that consists of **e**, together with all of its causes back to some arbitrary earlier time t. Then any possible structure of events that exists in a world with the same laws, and that has the same intrinsic character as S, *also* has the same causal character, at least with respect to the causal generation of **e**.

For the purposes of this section, we can read "has the same intrinsic character as" as "perfectly duplicates"—we won't need to compare structures of events that exhibit near-but-not-quite-perfect match of intrinsic character.

Now for some more detail. When Billy shot him down, Enemy was waiting for his home base—hundreds of miles away—to radio him instructions. At that moment, Enemy had no particular intention of going after Suzy; he was just minding his own business. Still, if Billy hadn't pulled the trigger, then Enemy would have eluded him, and moments later would have received instructions to shoot down the nearest suitable target (Suzy, as it happens). He would then have done so. But Billy does shoot him down, so he never receives the instructions. In fact, the home base doesn't even bother to send them, since it has been monitoring Enemy's transmissions and knows that he has been shot down.

Focus on the causal history of the bombing, back to the time of Billy's action. There is, of course, the process consisting of Suzy flying her plane, etc. (and, less conspicuously, the process consisting in the persistence of the target). If Dependence is true, then the causal history must also include Billy's action and its immediate effects: the bullets flying out of his gun, their impact with Enemy's fuselage, the subsequent explosion. (Perhaps we should also throw in some omissions: the failure of Enemy to do what he would have done, had he somehow eluded Billy. It makes no difference, since their contribution to the *intrinsic character* of the resulting causal history is nil.) Let this structure of events be S.

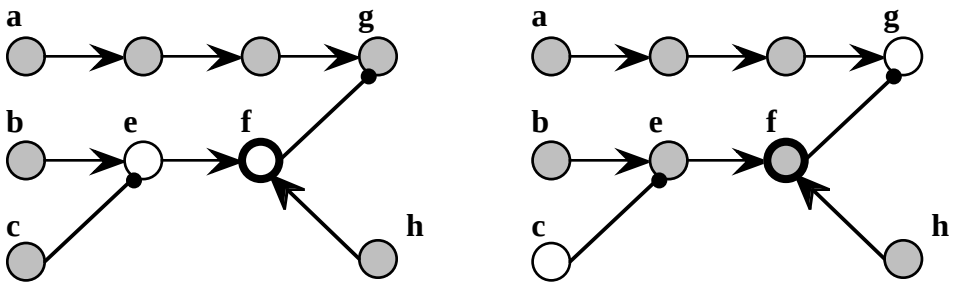
Two problems now emerge. In the first place, the intrinsic character of S fails to determine, together with the laws, that there are no *other* factors that would (i) stop Enemy, if Billy somehow failed to; (ii) do so in a way which would reverse the intuitive verdict (such as it is) that Billy's action is a cause of the bombing. Suppose, for instance, that we change the example by adding a bomb under Enemy's seat, which would have gone off seconds after the time at which Billy fired. And suppose that within this changed example, we can find a duplicate of S—in which case the specification of the intrinsic character of S must leave out the presence of the bomb. That shows (what was, perhaps, apparent already) that the dependence of the bombing on Billy's action is a fact *extrinsic* to S. If we decide that in this changed example, Billy's action is *not* a cause of the bombing (since, thanks to the bomb under Enemy's seat, he in fact poses no threat to Suzy), then we must either give up the Intrinsicness thesis, or grant that the causal history of the bombing (back to the time of Billy's action) wasn't described completely by S. Neither option is attractive. Let us call this the problem of the *extrinsic absence of disabling factors* (disabling in the sense that if they were present, there would be no dependence of the bombing on Billy's action).

Much more serious is the problem of the extrinsic *presence* of *enabling* factors (enabling in the sense that if they were absent, there would be no dependence of the bombing on Billy's action). For consider a third case, exactly like the first except in the following critical respect: The home base has no intentions of sending Enemy orders to shoot anyone down. In fact, if Billy hadn't pulled the trigger, then the instructions from the home base would have been for Enemy to return immediately. So Enemy poses no threat whatsoever to Suzy. Hence Billy's action is *not* a cause of the bombing. Yet the structure of events S is duplicated *exactly* in this scenario. So if the Intrinsicness thesis is right, then that causal history S must not in fact have been *complete*; we must have mistakenly excluded some events for which the third scenario contains no duplicates. Presumably, these events will be the ones that constitute the monitoring of Enemy by his home base, together with the intentions of his superiors to order him to shoot down the nearest appropriate target.

But now we are forced to say that these events count as *causes* of the bombing. That is ridiculous. It is not that they have *no* connection to the bombing, it's just that their connection is much more oblique:

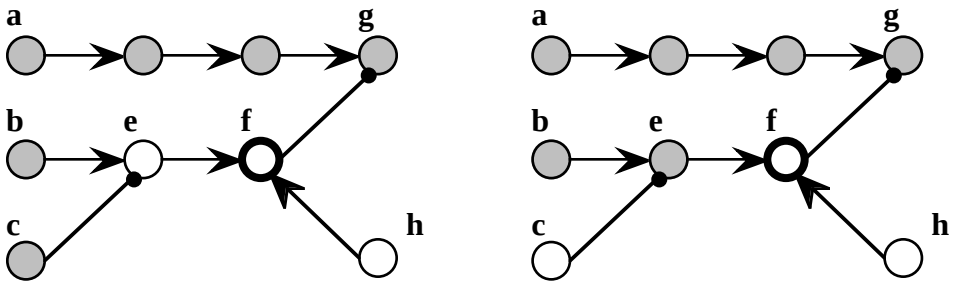
all we can say is that if they hadn't happened, then the bombing would not have depended on Billy's action. And notice, finally, that it is exactly the inclusion of Billy's action as part of the causal history S that is the culprit: Once we include it, we must also include (on pain of denying Intrinsicness) all those events whose occurrence is required to secure the counterfactual dependence of the bombing on this action.

To see this problem more vividly, compare the events depicted in figures 5 and 6:



c fires if **c** hadn't fired…

Figure 5



c fires if **c** hadn't fired…

Figure 6

Here, **f** is a stubborn neuron, needing two stimulatory signals in order to fire. Neuron **h**, in figure 5, fires shortly after the time at which neurons **a**, **b**, and **c** all fire (so I have abused the usual left-to-right conventions slightly). In the left-hand diagram of figure 5, **g** depends on **c**, but in figure 6 it does not; indeed, it would be quite ridiculous to claim, about the left-hand diagram of figure 6, that **c** was in *any* sense a cause of **g**.

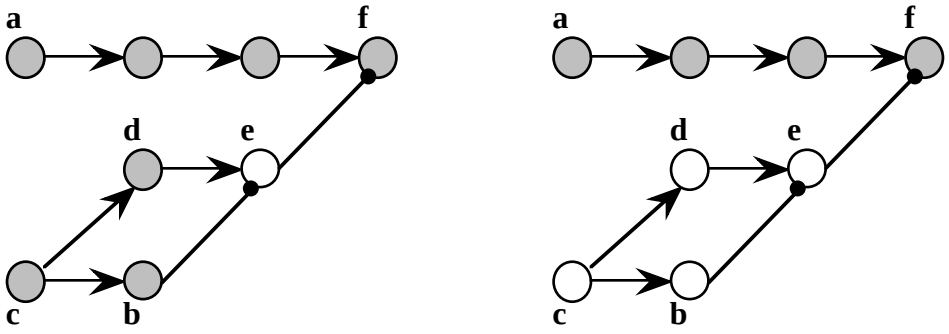
But now consider the causal history of **g**, in the left-hand diagram of figure 5, and suppose that—in keeping with Dependence—we count **c** as part of this causal history. Then it would seem that this causal history is duplicated *exactly* in the left-hand diagram of figure 6—in which case either Intrinsicness is false, or **c** in figure 6 is, after all, a cause of **g**. The only way out of this dilemma is to deny Dependence—or else to insist, against all good sense, that the causal history of **g**, in figure 5, *also* includes the firing of **h** (which is not duplicated in figure 6). But of course it does not: in figure 5, the firing of **h** is necessary, in order for **g** to depend on **c**; but that does not make it one of **g**'s causes.

§4.4Problems with transitivity

A more striking problem appears when we focus on the transitivity of causation. I begin by adding yet more detail to the example.

Early in the morning on the day of the bombing, Enemy's alarm clock goes off. A good thing, too: if it hadn't, he never would have woken up in time to go on his patrolling mission. Indeed, if his alarm clock hadn't gone off, Enemy would have been nowhere near the scene at which he was shot down. It follows that if Enemy's alarm clock hadn't gone off, then Billy would not have pulled the trigger. But it is also true that if Billy hadn't pulled the trigger, then the bombing would never have taken place. By transitivity, this ringing is one of the causes of the bombing.

Figure 7 helps to reinforce the absurdity of this conclusion:



c fires if **c** hadn't fired…

Figure 7

Neuron **e** can never fire. If **c** does not fire, then **e** won't get stimulated by **d**, whereas if **c** *does* fire, then the stimulation from **d** will be blocked by the inhibitory signal from **b**. So **e** poses no threat whatsoever to the firing of **f**. The little four-neuron network which culminates in **e** is, from the standpoint of **f**, totally inert.

Clearly, **c**'s firing cannot be a cause of **f**'s firing. At most, we might characterize **c**'s firing as something which *threatens to prevent* **f**'s firing, by way of the **c**-**d**-**e** connection—with the threat blocked by the **c**-**b**-**e** connection. Yet if both Dependence and Transitivity are correct, then **c**'s firing *is* a cause of **f**'s firing. For if **c** hadn't fired, then **b** would not have fired. Likewise, if **b** had not fired, then **f** would not have fired (recall here that backtracking is forbidden: we cannot say that if **b** had not fired, then it would have been that **c** didn't fire, and so **f** would have fired all the same). Since **f** depends on **b**, and **b** depends on **c**, it follows from Dependence and Transitivity that **c**'s firing is a cause of **f**'s firing. That consequence is unacceptable.

Certain examples with this structure border on the comic. Billy spies Suzy about to throw a rock at a window. He rushes to stop her, knowing that as usual he's going to take the blame for her act of vandalism. Unfortunately for him, trips over a tree-root, and Suzy, quite oblivious to his presence, goes ahead and breaks the window. If he hadn't tripped, he would have stopped her—so the breaking depends on the tripping. But if he hadn't *set out* to stop her, he wouldn't have tripped—so, by the combination of Transitivity and Dependence, he has helped cause the breaking after all, merely by setting out to stop it! That conclusion is, of course, just silly.[18]

Conclusion: If the thesis of Dependence is true, then each of Locality, Intrinsicness, and Transitivity is false. More precisely, if Dependence is true at a world, and the events in that world exhibit a causal structure rich enough to provide even one case of double-prevention like each of the ones we have been examining, then each of Locality, Intrinsicness, and Transitivity is false at that world. In the next section, we'll see that an exactly parallel conclusion can be drawn with respect to the thesis that omissions can be causes and effect.

§5 Omissions

The thesis of Omissions brings in its wake a number of difficult questions of ontology: Does it imply a commitment to a peculiar kind of "event" whose occurrence conditions essentially involve the *failure* of some ordinary type of event to occur? Does it make sense to speak of "the failure of **c** to occur", where "**c**" is supposed to refer to some ordinary event? (For perhaps such singular reference to non-actual events is impossible; alternatively, perhaps it is possible, but the circumstances in which we want to cite some omission as a cause or effect typically underdetermine which ordinary event is "omitted".) Do omissions have locations in space and time? If so, what determines these locations? (Recall the remarks in §4.2: right now I am at home, and hence fail to be in my office; is this omission located there or here?) And so on. I am simply going to gloss over all of these issues, and assume that a counterfactual supposition of the form "omission **o** does not occur" is equivalent to the supposition that some ordinary event of a given type C *does* occur (at, perhaps, a specific place and time)—where the type in question will be fixed, somehow, by the specification of **o** (or perhaps by context, or perhaps by both). At any rate, however justified complaints about the ontological status of omissions might be, they are emphatically not what is at issue, as we're about to see.

In what follows, I'll make the case that examples of causation *by* omission routinely violate each of Locality and Intrinsicness. The techniques I employ can be adapted so straightforwardly to make the same points about prevention (i.e., causation *of* omission) that we can safely leave those cases aside. Displaying the conflict between Omissions and Transitivity will require a case in which we treat an omission as an effect of one event and as a cause of another.

Finally, I am also going to gloss over the remarkably tricky question of when, exactly, we *have* a case of causation by or of omission—a question to which the thesis of Omissions only gives the vague answer, "sometimes." For example, is it enough to have causation of **e** by the failure of an event of type C to occur for **e** to counterfactually depend on this omission? Or must further constraints be satisfied? If not—if dependence is all that is required—we get such unwelcome results as that my act of typing has among its causes a quite astonishing multitude of omissions: the failure of a meteorite to strike our house moments ago, the failure of the President to walk in and interrupt me, etc. If, on the other hand, we insist that mere dependence is not enough for causation by omission, then we face the unenviable task of trying to characterize the further constraints. I'm going to sidestep these issues by picking cases that are uncontroversial examples of causation by omission—uncontroversial, that is, on the assumption that there are *any* such cases.

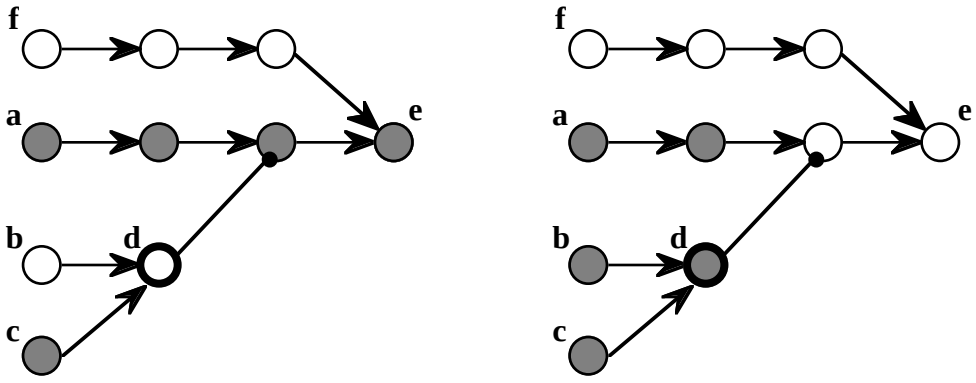
§5.1Problems with Locality

We can draw on the story of Suzy, Billy, and Enemy to show that, even if we waive worries about whether omissions have determinate locations, Locality fails for typical cases of causation by omission. Focus on a time t at which Enemy would have been approaching Suzy to shoot her down, had he not been shot down himself. Had Enemy not been absent, Suzy's mission would have failed; so the bombing depends on, as we might put it, the omission of Enemy's attack. More than this: The omission of Enemy's attack is among the *causes* of the bombing—at least, if there is to be causation by omission *at all*, this case should certainly be an example. But once again, it appears that the connection between this omission and the bombing must also qualify as a case of action at a distance, for no spatiotemporally continuous sequence of causal intermediates connects the two events. As before, the problem is not with finding a suitable location for the omission; it is rather that nothing guarantees that the sequence of omissions that proceeds from it (Enemy's failure to approach, pull the trigger, etc.) will intersect Suzy's *actual* flight. We can grant that the region of spacetime in which these omissions "take place" intersects the region she *would have* occupied, had Enemy not been absent. But it commits the same mistake as before to suppose that this region is the same as the region she *actually* occupies.

§5.2Problems with Intrinsicness

Whatever omissions are, they are notably lacking in intrinsic character. We already saw that for this reason, the Intrinsicness thesis needed to be phrased rather carefully: When we have picked out an event **e** and a structure of events S comprising **e** and all causes of **e** back to some earlier time, it is to be understood that any structure against which we compare S is composed solely of *genuine* events, not omissions. (On the other hand, no harm comes of letting *S* include omissions, at least on the assumption that they contribute nothing to its intrinsic character.) Still, it is for all that consistent to hold that Intrinsicness *applies* to causation by omission, as follows: Suppose that **e** occurs at time t', and that S consists of **e** and all causes of **e** back to some earlier time t. Suppose further that we count the omission **o** as one of **e**'s causes, and that **o** "occurs" (in whatever sense is appropriate for omissions) in the interval between t and t'. Then if structure S' intrinsically matches S, there must be some omission **o**' "corresponding" to **o** that causes the event **e**' in S' that corresponds to **e** in S (never mind that **o**' is not *part* of S'). In short, we might think that causation of an event by omission supervenes on the intrinsic character of that event's "positive" causal history.

This conjecture is false. To show why, I'll argue that both of the problems we saw in §4.3—the problem of the extrinsic lack of disabling factors and the problem of the extrinsic presence of enabling factors—recur in this context. A simple neuron diagram will serve to illustrate each:



b doesn't fire if **b** had fired…

Figure 8

In figure 8, **d** is a dull neuron that needs two stimulatory signals in order to fire. Thus, **d** fails to fire even though stimulated by **c**; still, since **c** fires, **e**'s firing depends on the *failure* of **b** to fire (at, say, time t, which we will take to be the time of **a**'s firing). Note finally that if **b** had fired and **f** had as well (at t), then **e** would have fired all the same.

Let us suppose, in keeping with the Omissions thesis, that the failure of **b** to fire at t is among the causes of **e**'s firing. Let S consist of **e**, together with all of its (positive) causes back to time t. Then if Intrinsicness applies to causation by omission in the way we have suggested, any nomologically possible structure that duplicates S will exhibit the same causal relationships: in particular, there will be an omission that "duplicates" **b**'s failure to fire and that will be a cause of the event that duplicates **e**'s firing. Here is one such possible structure, embedded in slightly different surroundings:

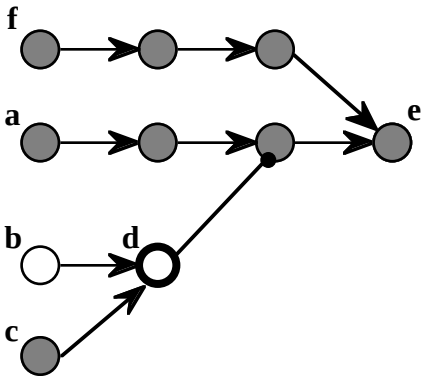


Figure 9

And here is another, again in different surroundings:

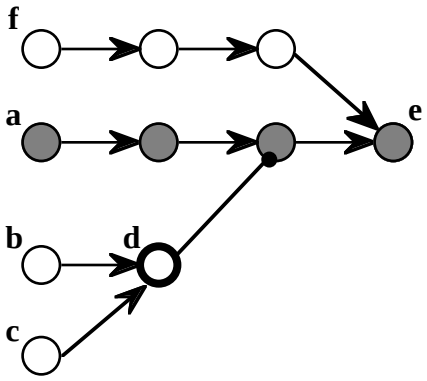


Figure 10

The problem is that in each case, **b**'s failure to fire is no longer a cause of **e**'s firing, *contra* the requirements of our conjecture about how Intrinsicness covers causation by omission. In figure 9, the firing of **f** renders **b**'s failure to fire quite irrelevant to whether **e** fires, showing that when **b**'s failure to fire *is* a cause of **e**'s firing, this is due in part to the extrinsic absence of disabling factors. Likewise, in figure 10, **c**'s failure to fire renders the behavior of **b** irrelevant, showing that when **b**'s failure to fire is a cause of **e**'s firing, this is due in part to the extrinsic presence of enabling factors. So the leading idea behind the Intrinsicness thesis—namely, that it is the intrinsic character of some event's causal history that (together with the laws) makes it the case that this *is* its causal history—comes directly into conflict with the Omissions thesis.

§5.3Problems with Transitivity

As before, more striking problems emerge when we combine the theses of Omissions and Transitivity. To see how easy it is to concoct an absurdity from these two ingredients, consider the following variant on our story: This time, Enemy's superiors on the ground had no intention of going after Suzy—until, that is, Billy shoots Enemy down. Outraged by this unprovoked act of aggression, they send out an all-points-bulletin, instructing any available fighter to go after Suzy (a much more valuable target than Billy). Alas, Enemy was the only fighter in the area. Had he somehow been present at the time of the broadcast, he would have received it, and promptly targeted and shot down Suzy; his *absence* is thereby a cause of the bombing. But, of course, his absence is itself caused by Billy's action. So by Transitivity, we get the result that Billy's action is a cause of the bombing. Lest the details of the case be distracting, let's be clear: *all Billy does it to provoke a threat to the bombing;* luckily for him, the very action that provokes the threat also manages to counteract it. Note the similarity to our earlier "counterexample" to Transitivity: Enemy's action (taking off in the morning) both causes a threat to the bombing (by putting Enemy within striking range of Suzy) and counteracts that threat (by likewise putting Enemy within Billy's striking range).

Conclusion: If the thesis of Omissions is true, then each of Locality, Intrinsicness, and Transitivity is false. More precisely, if Omissions is true at a world, and the events in that world exhibit a causal structure rich enough to provide cases of the kinds we have just considered, then each of Locality, Intrinsicness, and Transitivity is false at that world.

§6Diagnosis: two concepts of causation

Here are two opposed reactions one might have to the discussion so far:

Counterfactual dependence is *not* causation. In the first place, it's not (as everyone recognizes) necessary for causation. In the second place, the best attempts to tart it up in such a way as to yield a full-blown analysis of causation rely on the three theses of Locality, Intrinsicness, and Transitivity—and the lesson of double prevention (a lesson also supported by considering the causal status of omissions) is that these theses *contradict* the claim that dependence is sufficient for causation. The theses are too important; this latter claim must be given up. But give up Dependence, and you've torn the heart out of counterfactual analyses of causation.

Nonsense; counterfactual dependence *is too* causation. Here we have two wholly distinct events; moreover, if the first had not happened, then the second would not have happened. So we can say—notice

how smoothly the words glide off the tongue!—that it is in part *because* the first happened that the second happened, that the first event is partly *responsible for* the second event, that the occurrence of the first event helps to *explain why* the second event happened, etc. Nor do we reverse these verdicts when we discover that the dependence arises by way of double prevention; that seems quite irrelevant. All of these locutions are *causal* locutions, and their appropriateness can, quite clearly, be justified by the claim that the second event counterfactually depends on the first event. So how could this relation fail to be causal? To be sure, it's another question whether we can use it to construct a full-blown analysis of causation, but at the very least we have the result that counterfactual dependence (between wholly distinct events) is *sufficient* for causation—which is just to say that Dependence *is* true.

The claims of both of the foregoing paragraphs are correct. But not by making a contradiction true: rather, what is meant by "causation" in each case is *different*. Counterfactual dependence is causation in one sense: but in *that* sense of "cause", Transitivity, Locality, and Intrinsicness are all false. Still, they are not false *simpliciter*; for there is a *different* concept of causation—the one I call "production"—which renders them true. Thus, what we have in the standard cases of overdetermination we reviewed in §3 are not merely counterexamples to some hopeless attempt at an analysis of causation, but cases that reveal one way the concepts of dependence and production can come apart: These cases uniformly exhibit production without dependence. What we have in the cases of double-prevention and causation by omission we examined in §§4-5 are not merely more nails in the coffin of the counterfactual analysis, but cases that reveal the other way the two causal concepts can come apart: For these cases uniformly exhibit dependence without production. Similarly, we can now diagnose the intuitions Bennett is pumping in his April rains/June forest fire case. For while there is a sense in which the rains *do* cause the fire—the fire clearly depends on the rains—there is an equally good sense in which they don't—the rains do not help to produce the fire. That is because (surprise!) we have here a case of double-prevention: The rains prevent an event (fire in May) which, had it occurred, would have prevented the June fires (by destroying the flammable material).

The principle virtues of my claim are thus clear: It allows us to maintain each of the five theses. It provides us with a natural and compelling diagnosis of the most important problem cases for analyses of causation. And it should come as no surprise that the distinction between production and dependence has gone unnoticed, for *typically* the two relations coincide (more exactly, I think, production typically coincides with the ancestral of dependence; more on this in §7.4, below).

An additional virtue of the position, perhaps less obvious than the foregoing ones, concerns the ontological status of omissions. Those who endorse the Omissions thesis might worry that they are thereby committed to the existence of a special sort of event—as if the truth of "the failure of an event of type C to occur caused **e** to occur" required the existence of something that answered to the description, "failure of an event of type C to occur." But if the only sense in which omissions can cause and be caused is that they can enter into relations of counterfactual dependence, then this worry is quite misplaced. For talk of causation by and of omissions turns out to be nothing more than a way of talking about claims of the form, "if an event of type C had occurred, then..." and "if ..., then an event of type C would have occurred." Manifestly, neither locution carries an ontological commitment to a strange sort of "negative" event. So, if I am right, anxieties about whether we can find a place for omissions in the causal order rest on a basic confusion about what it means to attribute causal status to omissions.

This observation connects to a broader point, which is that dependence, understood as a relation *between events*, is unduly restrictive. For quite generally there can be counterfactual dependence between *facts* (true propositions), where these can be "positive", "negative", "disjunctive", or whatever—and where only rarely can we shoehorn the facts so related into the form, "such-and-such an event occurred". When we can—when we can say that the fact that **e** occurred depends on the fact that **c** occurred—then we can go ahead and call this a kind of event-causation. But to see it as anything but a special case of a causal relation with a much broader domain would be, I think, a mistake.[19]

We can bring my thesis into still sharper focus by considering the some of the more obvious objections to it. It seems wise to begin by directly confronting what many will see as the most damning objection—which is simply that it posits *two* concepts of event-causation! This might strike some as an extravagantly high price to pay: After all, when possible we should be conservative, and conservatism argues for taking our concept of event-causation to be univocal. At the very least, shouldn't we view the bifurcation of our concept of event-causation as a very serious cost of my proposal?

No, we should not—and not because we shouldn't be conservative. It's rather that this objection mistakes a perfectly sensible *methodological maxim* with a *reason to believe*. The methodological maxim goes: When trying to come up with an analysis of a concept, start out by operating under the assumption that the concept is univocal. I think that's sound. But it doesn't at all follow that it is somehow

antecedently more probable that the concept in question is univocal—let alone *so* probable that any analysis that says otherwise pays a "high price". In the face of the right sorts of reasons to prefer a non-univocal analysis, we should give up our operative assumption—and we shouldn't expect those reasons to have to carry an extra heavy burden of proof because of the "intrinsic plausibility" of the hypothesis of univocality.

To think otherwise manifests a basic confusion. It's rather as if I had lost my keys somewhere in this room; I have no idea where. They might be over there, where it's dark and a lot of debris obscures things; or they might be over here, where it's sunny and uncluttered. It makes exceedingly good sense for me to start by looking in the sunny and uncluttered part of the room—to *act as if* I believed my keys were there. But that is not because I *do* believe they are there, or even because I consider it more likely than the alternative (as if the hypothesis that life is easy has some intrinsic plausibility to it!). It's rather that if my keys *are* in the uncluttered area, then I will soon find them—and if they are not, I will quickly find that out as well.

In the same way, when we go to analyze some concept of philosophical interest, it makes exceedingly good sense to start by looking for a univocal analysis. For even if we are wrong, and some hidden ambiguity lurks in our ordinary applications of the concept, the very problems we will encounter in trying to come up with a univocal analysis will (if we are careful and attentive) be diagnostic of this ambiguity. (The critique of the counterfactual analysis carried out in §§3-5 was partly designed to be a case in point.) But it is foolishness to mistake this advice for a reason to *believe* that the concept is univocal. Indeed, if I consider the hypothesis that our concept of event-causation is univocal, I see no reason whatsoever to judge it to be highly probable, antecedently to any investigation. And *after* sufficient investigation—in particular, after very basic principles governing our application of "cause" have been shown to come into conflict—I think its plausibility is just about nil.

A more subtle objection is the following: What I have really shown is not that there are two *concepts* of causation, but rather that there are two *kinds* of causation, two different ways in which one event can be a cause of another. That may well be right; certainly, I was happy to begin this paper by announcing that event-causation comes in two "varieties". I do not know how to judge the matter, because I am not sufficiently clear on what underlies this distinction between concepts and kinds. Compare a nice example borrowed from Tim Maudlin: There are at least three different ways of being a mother. We might call them "DNA-mother", "womb-mother", and "nurturing mother". Does that mean we have three different concepts of mother—an ambiguity largely unnoticed only because those we call "mothers" are typically all three? I don't know. At any rate, in the case at hand it doesn't matter in the slightest. I am quite content to agree that I have (merely) shown that there are two kinds of causation—as long as those who insist on this rendering of my thesis agree that the two kinds answer to very different criteria, and consequently require very different analyses. That claim alone is enough to show how unwise it would be, when attempting to provide a philosophical account of event-causation, merely to forge blindly ahead, trying to come up with an analysis which can successfully run the gauntlet of known problem cases. If I am right, any such *single* analysis is doomed to failure.

A third, more congenial objection begins by granting the distinction between production and dependence, but denying that dependence deserves to be counted a kind of causation at all. Now, I think there is *something* right about this objection, in that production does seem, in some sense, to be the more "central" causal notion. As evidence, consider that when presented with a paradigm case of production without dependence—as in, say, the story of Suzy, Billy, and the broken bottle—we unhesitatingly classify the producer as a cause; whereas when presented with a paradigm case of dependence without production—as in, say, the story of Suzy, Billy, and Enemy—our intuitions (well, those of some of us, anyway) about whether a genuine causal relation is manifested are shakier. Fair enough. But I think it goes too far to deny that counterfactual dependence between wholly distinct events is not a kind of *causal* relation. Partly this is because dependence plays the appropriate sort of roles in, for example, explanation and decision. (See §8, below, for more discussion of this point.) And partly it is because I do not see how to accommodate causation of and by omissions (as we should) as a species of production; counterfactual dependence seems the only appropriate causal relation for such "negative events" to stand in.

This last point brings up a fourth possible objection, which is that in claiming that there are two kinds of causation, each characterized by a different subset of the five theses, I have overstepped my bounds. After all, even if the arguments of §§4-5 succeed, all they establish is, roughly, (i) that Dependence contradicts each of Locality, Intrinsicness, and Transitivity; and (ii) that Omissions likewise contradicts each of Locality, Intrinsicness, and Transitivity. It obviously doesn't follow that Dependence and Omissions should be bundled together and taken to characterize one kind of causation, nor that Locality,

Intrinsicness, and Transitivity should be bundled together and taken to characterize another. Perhaps the ambiguity in our ordinary causal talk is more multifarious and messy than this claim allows.

Dead right. And even though I think that further investigation could unearth more positive reasons for dividing the five theses into the two groups I have chosen, I do not have such reasons to offer here. For what it's worth, I do have a strong hunch that, as noted above, there couldn't *be* anything more to causation of and by omissions than counterfactual dependence; hence the pairing of Omissions with Dependence. And in the next section I'll propose an analysis of production that gives central roles to both Intrinsicness and Transitivity, as well as to a slightly weakened version of Locality. But that's hardly enough to warrant conviction. Rather, what's wanted are more probing arguments as to why our ordinary notion of event-causation should fracture cleanly along the lines I have drawn. Lacking such arguments, I will fall back on the methodological maxim discussed above: Given that we can no longer take it as a working hypothesis that the concept of causation is univocal, let us nevertheless adopt the most conservative working hypothesis available to us. Since we have yet to find any reason to think that Dependence conflicts with Omissions, or that conceptual tensions threaten the happy union of Locality, Intrinsicness, and Transitivity, let us assume—again, as a working hypothesis—that the first two theses characterize one causal notion, the last three another.

And let us now consider how the two causal notions are to be analyzed.

§7 The two concepts analyzed

Part of the task—analyzing dependence—is easy: it is simply counterfactual dependence between distinct events. More cautiously, we might want to admit another kind of counterfactual dependence as well. Perhaps counterfactual *covariation*—manifested when the time and manner of one event's occurrence systematically counterfactually depend on the time and manner of another's—should count as a kind of causation as well, to be classified as a close relative of dependence. (It's clearly possible to have dependence without covariation, as in typical cases of double-prevention; Schaffer (2000) provides compelling examples of covariation without dependence, as well.) No matter; given that these counterfactual locutions are themselves well-understood, our work here is basically already done for us. (But see §8.2 for some tentative reservations.)

Production is harder. In this section I will put forth my own proposal—speculative, and, as we'll see, somewhat limited in scope—for a reductive analysis of this relation. I will set it out in two parts. The first, less speculative part outlines a certain strategy for developing an analysis, which I call the "blueprint strategy". The second, more speculative part describes my (currently) preferred way of implementing this strategy.

§7.1 The blueprint strategy

Suppose we have an analysis that succeeds—when circumstances are nice—in singling out a portion of the causal history of some target event **e**, where this is understood to be the history of **e**'s *producers*. (When circumstances are not nice, let the analysis fall silent.) It might be a simple counterfactual analysis: When circumstances are nice (when there is no double-prevention, or overdetermination…), the causal history of **e** back to some earlier time t consists of all those events occurring in that interval upon which **e** depends. Or it might be a Mackie-style analysis: the causal history consists of all those events (again, occurring in that interval) which are necessary parts of some sufficient condition for **e**. Or it might be some other kind of analysis. Then—provided we can say with enough precision what it takes for circumstances to be "nice"—we can use the Intrinsicness and Transitivity theses to extend the reach of this analysis, as follows:

First, suppose we examine some events **c** and **e**, and find that our analysis is silent as to whether **c** is a cause of **e**. Still, we find that **c** and **e** belong to a structure of events S such that (i) S intrinsically matches some *other* structure of events S' (occurring in a world with the same laws as the world of S); and (ii) our analysis counts S' as a segment of the causal history of **e**' (where **e**' is the event in S' which corresponds to **e** in S). That is, our analysis counts S' as a *rich enough set of causes of **e**' for the Intrinsicness thesis to apply*. It follows that S has the same causal structure as S' (at least, with respect to the target event **e**), hence that **c** is a cause of **e**.

For convenience, let us say that when the conjunction of our analysis with the Intrinsicness thesis counts **c** as a cause of **e**, **c** is a "proximate cause" of **e**. Then, second, we parlay proximate causation into causation *simpliciter* by means of the Transitivity thesis: causation is simply the ancestral of proximate causation. In short, we use our original analysis to find a set of *blueprints* for causal structures, which we can then use to map out (if we are lucky) the causal structure of *any* set of events, in *any* circumstances, by means of the Intrinsicness and Transitivity theses.

This strategy has the virtue of factoring the analysis of production into two parts: the analysis which will produce the "blueprints", and the extension of any such analysis into a full analysis of production by means of the Intrinsicness and Transitivity theses. Still, two potential difficulties deserve mention. First, recall that the Intrinsicness thesis as I've stated it presupposes that there is neither action at a temporal distance nor backwards causation—so without a more general statement of the Intrinsicness thesis, the full analysis of production will necessarily be limited in scope. Second, recall that for the purposes to which I put the Intrinsicness thesis above (to reveal conflicts with Dependence and Omissions), the "same intrinsic character" clause in that thesis could be understood in a relatively clear and uncontroversial sense, namely as requiring that the two structures of events in question be *perfect duplicates*. That is, to make trouble for Dependence and Omissions we only needed to assume, roughly, that two event-structures that *perfectly* match one another in intrinsic respects likewise match in causal respects. But the blueprint strategy affords us no such luxury.

To see why, consider our old standby example of Billy, Suzy, and the broken bottle. Suppose that our unadorned analysis (whatever it turns out to be) falls silent about whether Suzy's throw is a cause of the breaking—and this, thanks to the confounding presence of Billy's throw. And suppose that the counterfactual situation in which Billy's throw is absent is one whose causal structure our analysis succeeds in capturing—in particular, the counterpart to Suzy's throw, in that situation, is counted a cause of the counterpart to the breaking. Victory! —For surely we can say that when Billy's throw is present, Suzy's still counts as a cause, because it belongs to a structure of events (the throw, the flight of the rock, etc.) that matches an appropriate "blueprint" structure—namely, the structure found in the counterfactual situation where Billy's throw is absent. Don't we have here a vindication of the blueprint strategy?

Yes, but only if the notion of "matching" is more liberal and, regrettably, vague than the restrictive, relatively precise notion of *perfect* match. For the two sequences of events—the one beginning with Suzy's throw, in the case where Billy also throws, and the one beginning with her throw, in the case where he doesn't—will *not* match perfectly: for example, tiny gravitational effects from Billy's rock will guarantee that the trajectories of Suzy's rock, in each case, are not *quite* the same. So we are left with the unfinished business of saying what imperfect match consists in, and of specifying how imperfect it can be, consistent with the requirements of the blueprint strategy. While I do not think these difficulties undermine the blueprint strategy, I won't try to resolve them here. (But see my "The intrinsic character of causation" for some suggestions.)

§7.2Implementing the strategy (first pass)

Now for my own story about what makes for "nice" circumstances, and how an analysis should proceed, under the assumption that they obtain. As usual, I will assume determinism, but I will also assume that there is no action at a temporal distance, nor backwards causation (not merely because I wish to slot the following analysis into the blueprint strategy, but also because I do not yet know how to make the analysis itself work, without these assumptions.) First, some terminology.

Suppose that at time t, the members of some set S of events all occur, and that **e** occurs at some later time t'. I will say that S is *sufficient* for **e** just in case the fact that **e** occurs follows from

(i) the laws, together with

(ii) the premise that all the members of S occur at t, together with

(iii) the premise that no other events occur at t.

The entailment here is metaphysical, not narrowly logical. I will also take as a requirement that the entailment not be vacuous: (i) – (iii) must be consistent. I will say that S is *minimally sufficient* for **e** just in case S is sufficient for **e**, but no proper subset of S is. (We might want to add a premise to the effect that relevant background conditions obtain. I prefer to treat any such conditions as "encoded" as members of S.) Do not be distracted by the fact that in typical situations, (iii) will be *false* (though (i) and (ii) will of course be true); that is quite irrelevant to the purposes to which we will put the notions of sufficiency and minimal sufficiency. Finally, the quantifier in (iii) must be understood as ranging over only *genuine* events, and not omissions, else the premise is inconsistent all by itself. Suppose, for example, that our set S does *not* include a kiss at a certain location l; the "no other events occur" requirement will therefore entail that no kiss occurs at t at l. Then consider the omission **o**, which consists in the *failure* of a kiss to take place at t at l, and suppose that **o** is also not a member of S. To require, in addition, that this "event" not occur, is just to require that a kiss *does* take place at t at l. Quite obviously, we can't add that requirement consistently.

Roughly, we can say that where the members of S occur at t, S is sufficient for later event **e** just in case, had only the events in S occurred at t, **e** would still have occurred; S is minimally sufficient if the same is not true for any proper subset of S. (Since I have employed a counterfactual locution here, one

might want to call the resulting analysis a counterfactual analysis! Call it what you will—just don't confuse it with those analyses that take Dependence as their starting point.)

It seems that the problems that confound the usual attempts to analyze causation all have to do with stuff going on in the environment of the genuine causal process, stuff that ruins what would otherwise be the neat nomological relationships between the constituents of that process and the given effect. An attractive and simple idea is that *if*, at a time, there is a unique minimally sufficient set for our target effect **e**, then such environmental "noise" must be absent—so that circumstances are "nice"—and we can take it that *this unique minimally sufficient set contains all and only the producers of **e** that occur at that time*. If so, then one way of implementing the blueprint strategy becomes obvious: Suppose that **e** occurs at t', and that t is an earlier time such that at each time between t and t', there is a unique minimally sufficient set for **e**; then the segment of **e**'s causal history back to time t consists of all and only the events in these sets.

It will turn out that this simple idea won't work without a significant adjustment. But first some good news.

One embarrassment for a Mackie-style analysis is the so-called problem of common effects (figure 11): **d** and **e** are both effects of **c**, hence from the fact that **d** occurs (at t, say), together with the laws, together with an appropriate specification of the circumstances, it follows that **e** occurs (at t', say). (The reasoning is something like the following: given the circumstances, **d** could only occur if **c** caused it; but in that case **e** must have occurred as well.)

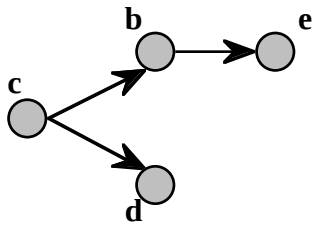
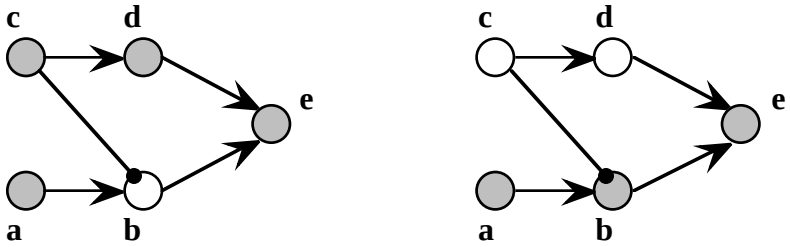


Figure 11

Our provisional analysis has no such problem, since **d** cannot be part of a minimally sufficient set for **e**. For suppose it is part of some such set S. **d** cannot be the only member, since it patently does not follow from the laws, together with the claim that **d** alone occurs at t, that **e** occurs at t'. So S must also include events whose occurrence consists in the presence, at t, of the appropriate stimulatory connections among the four neurons. (We'll assume that the laws, together with the fact that no other events occur, entail that if these connections are present at t, then they are also present a little bit before and a little bit after; without this assumption, there is clearly no hope of connecting **d**'s occurrence to **e**'s, via the laws.) But then it follows from the laws, together with the claim that the members of S occur at t, that **b** occurs at t as well; for there will be no way to secure the claim that **e** occurs at t' except by way of securing *this* claim. In that case, we face a dilemma: either **b** is a member of S—in which case **d** is redundant, and so S cannot be *minimally* sufficient—or **b** is not a member of S—in which case our three conditions (i) – (iii) are *inconsistent*, and so S is not sufficient.

So our analysis does not count **d** as a cause of **e**. Nor will it, when we add in the Intrinsicness and Transitivity theses: For **d** cannot inherit causal status from any blueprint; in order to do so the blueprint would have to contain a copy of **d**, a copy of **e**, and other events contemporaneous with the **d**-duplicate and with which it formed a minimally sufficient set. These other events, moreover, would have to be duplicated in the events of figure 11. In short, the duplicate would have to be the result of *subtracting* events from figure 11 in such a way that **d** remained, and belonged to a minimally sufficient set for **e**. But there is manifestly no way to perform such a subtraction. And it is equally clear that no *sequence* of blueprints can connect up **d** with **e**.

Next, consider ordinary preemption (figure 2):



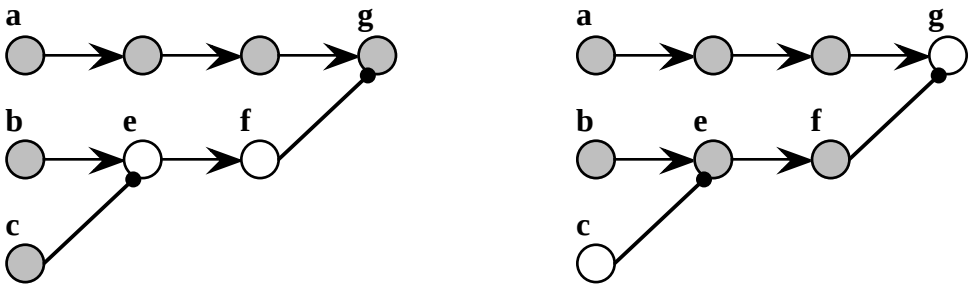
c fires if **c** hadn't fired…

Figure 2

At the time of occurrence of **a** and **c**, there are two minimally sufficient sets for **e**: {**c**} and {**a**}. (For ease of exposition, here and in the rest of this section I'll suppress mention of those events whose occurrence consists in the presence of the relevant stimulatory and inhibitory connections.) So circumstances are not "nice". Still, there is no problem with getting **c** to come out as a cause of **e**, for there is an obvious blueprint contained within the circumstances that would have occurred, if **a** had not fired. But no such blueprint connecting **a** with **e** can be found in the circumstances that would have obtained, had **c** not fired; for in those circumstances, the causal history of **e** will include the firing of **b**, and so there will be no "match" between this causal history and any part of the actual structure of events.

Next, late preemption: Happily, such cases receive exactly the same diagnosis as the case of ordinary preemption, and so need no special treatment.

Finally, double-prevention (figure 4):



c fires if **c** hadn't fired…

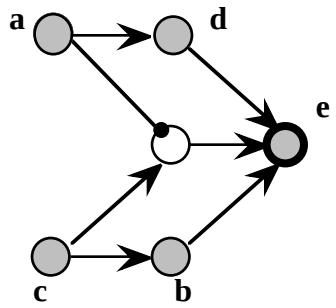
Figure 4

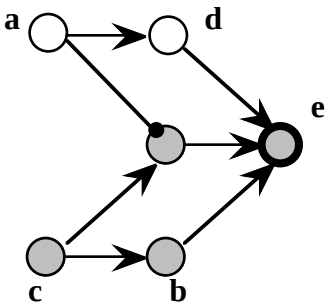
Observe first that there is, at the time of occurrence of **a**, **b**, and **c**, a unique minimally sufficient set for **g**: namely, {**a**}. So if **c** is to qualify as a cause of **g**, we must find a blueprint, or sequence of blueprints, which will connect up **c** with **g**. But that is evidently impossible. For such a blueprint would have to describe a causal history for **g** that is different from the one that actually obtains; otherwise, this causal history would contain a duplicate of **a**, in which case the duplicate of **c** could not be part of a minimally sufficient set for the duplicate of **g**. But it is apparent that there is no sequence of events connecting **c** with **g** that could serve as such an alternate causal history.

§7.3Implementing the strategy (final pass)

So far, so good. Unfortunately, two difficulties scotch the key idea that when there is a unique minimally sufficient set for **e** at a time, then its elements are all and only the producers of **e** (at that time).

First, there are producers that belong to no minimally sufficient set (figure 12):



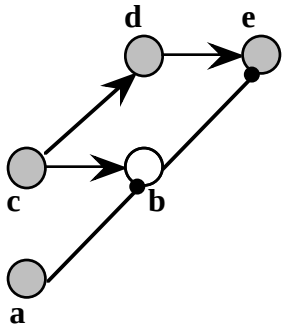


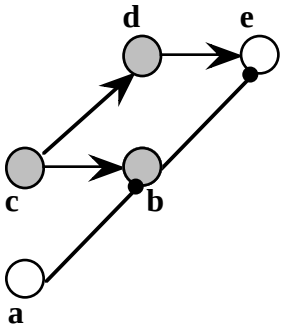
a fires if **a** hadn't fired…

Figure 12

In the diagram, **a** and **c** are clearly the producers of **e**; yet the unique minimally sufficient set for **e** contains just **c**. (Remember that **e**, here, is a stubborn neuron, requiring two stimulatory signals to fire.)

Second, there are non-producers that belong to unique minimally sufficient sets (figure 13):





a fires if **a** hadn't fired…

Figure 13

In the diagram, **a** is clearly not a producer of **e**; yet it is included in {**a**, **c**} the unique minimally sufficient set for **e**. (Notice that this is a special case of double-prevention, one that eludes the treatment we gave of the standard sort of case exhibited by figure 4.)

I suggest these problems arise because we have overlooked an important constraint on the internal structure of causal histories. Suppose that **e** occurs at time t_2, and that we have identified the set of its producers at both time t_0 (call this set S_0) and time t_1 (call this set S_1) ($t_0 < t_1 < t_2$). Then it had better be the case that when we trace the causal histories of the elements of S_1 back to t_0,

(i) we find no events *outside* of S_0—for otherwise transitivity of production would have been violated; and

(ii) we *do* find all the events *inside* S_0—for otherwise we would have to say that one of these events helped produce **e**, but not by way of any of the t_1-intermediates.

Return to the diagrams. In figure 12, our analysis tells us (among other things) that **d** is a producer of **e** and that **a** is a producer of **d**—but fails to deliver the consequence that **a** is a producer of **e**. That is an example of a failure to meet constraint (i). In figure 13, our analysis identifies **d** as the sole intermediate producer of **e**, and identifies **c** as *its* only producer—thus misdescribing **a** as a producer of **e** that somehow fails to act by way of any intermediates. That is an example of a failure to meet constraint (ii).

The way to fix these problem is to make our "nice circumstances" analysis still more restrictive, by building into it the two foregoing constraints. We begin as before, by supposing that **e** occurs at t', and that t is an earlier time such that at each time between t and t', there is a unique minimally sufficient set

for **e**. But now we add the requirement that whenever t_0 and t_1 are two such times ($t_0 < t_1$) and S_0 and S_1 the corresponding minimally sufficient sets, then

(i) for each element of S_1, there is at t_0 a unique minimally sufficient set; and

(ii) the union of these minimally sufficient sets is S_0.

This added requirement gives expression to the idea that when we, as it were, identify the producers of **e** *directly*, by appeal to their nomological relationship to **e**, we must get the same result as when we identify them by "tracing back" through intermediate producers.

We're now in a position to state the analysis of production:

Given some event **e** occurring at time t' and given some earlier time t, we will say that **e** has a *pure causal history* back to time t just in case there is, at every time between t and t', a unique minimally sufficient set for **e**, and the collection of these sets meets the two foregoing constraints. We will call the structure consisting of the members of these sets the "pure causal history" of **e**, back to time t.

We will say that **c** is a *proximate cause* of **e** just in case **c** and **e** belong to some structure of events S for which there is at least one nomologically possible structure S' such that (i) S' intrinsically matches S; and (ii) S' consists of an **e**-duplicate, together with a pure causal history of this **e**-duplicate back to some earlier time. (In easy cases, S will itself be the needed duplicate structure.)

Production, finally, is defined as the ancestral of proximate causation.

§7.4Refinements and observations

The structure of the analysis is easiest to understand if we make certain rather drastic simplifying assumptions about the nature of events. (In fact, we'll see that relaxing them will require some technical adjustments to the analysis.) Let us suppose that events are

(i) *momentary*, occurring, in their entirety, at single instants of time;

(ii) mereologically *atomic*, having no other events as parts; and

(iii) *modally precise*, in the sense that no event could occur at a time or in a manner other than it actually does. It follows from (iii) that

(iv) events are never distinguished merely by their modal properties.

Example: Suzy gives Billy a passionate kiss. Some would say that there are (at least) two events here, both kisses and both passionate, but distinguished in that one of them is *essentially* passionate (it could not have occurred without being passionate), whereas the other is only accidentally so. That would make (iv) false, and therefore also (iii), since (iii) denies that the accidentally passionate kiss exists.[20]

Now suppose we have fixed on some event **e** that occurs at t', and are looking to an earlier time t to find a unique minimally sufficient set for **e**. Any candidate set of events S will determine a *unique* physically possible state that the world might have had at t: namely, the state it would have had if the events in S had been the only events occurring at that time. What guarantees that this state is unique is simply that the events in S are modally precise: to say that they occur is automatically to specify the exact manner in which they occur, and thereby to specify exactly the physical state that obtains when they and no other events occur.[21] Then S is sufficient for **e** just in case this physical state, were it to obtain at t, would evolve via the laws into a state at t' in which **e** occurs. (Once we understand sufficiency, we understand minimal sufficiency, unique minimal sufficiency, etc.)

It's best to view these simplifying assumptions not as yielding an approximation to the analysis of production, but rather as yielding an accurate analysis of production, restricted to the special case of events that meet them. For the multitude of events that don't meet them—either because they last for more than an instant, or because they have parts, or because they could have occurred at a time or in a manner other than they actually do—the analysis needs various technical adjustments. The guiding idea behind the core of the analysis—that for a given effect **e** and a given candidate set S of producers of **e**, we consider what would have happened if only the events in S had occurred—still remains in place. But in order to accommodate events with duration, we should broaden our focus to include not just sets of events that *occur* at the earlier time t, but also sets of events that *are occurring* at t. And in order to accommodate the fact that these events might be capable of occurring at different times, we should consider not the hypothesis that they alone are occurring at t, but the weaker hypothesis that at *some* time, they alone are occurring. Finally, we must take care to understand the definitions of both "sufficient" and "unique minimally sufficient set" in a somewhat technical sense, once we admit events with mereological structure, or events that differ from one another merely in their modal properties. As to "sufficient": When we say that a set S of events, all of which are occurring at t, is sufficient for later event **e** just in case it follows that **e** occurs, from the laws and the premise that at some time *only the events in S are occurring*, we must understand the italicized phrase to mean: no event wholly distinct from every event in S is occurring. For obviously, if **c** is occurring, then so are (at least some of) its parts, and so are other events distinguished from it merely by their modal properties. (If Suzy's essentially passionate kiss is

occurring, then so is her accidentally passionate one, if there be such.) As to "unique minimally sufficient set": If a set S is minimally sufficient for **e**, then often so will other sets be that are obtained from S by replacement of a member event **c** by its parts, or by another event distinguished from **c** merely by its modal properties. That does not threaten uniqueness, understood properly: rather, uniqueness should be understood to fail iff there are two minimally sufficient sets, where at least one member of one is wholly distinct from every member of the other.

Investigating and developing these and no doubt other refinements is a worthy project, for another occasion. The analysis as stated is clear enough that we can make several useful observations about it.

First, the notion of production that the analysis characterizes quite obviously vindicates the Intrinsicness and Transitivity theses, since they are built right into it. Still, one might wonder whether they need to be built in so explicitly. Could the notion that producers are, roughly, members of unique minimally sufficient sets just yield one or the other of these theses as corollaries?

That would be nifty, but I think that it's too much to hope for. As to transitivity, consider again the events depicted in figure 13:

a fires if **a** hadn't fired…

Figure 13

Let us interpret the arrows in this diagram not as representing stimulatory connections between neurons, but as representing processes—real or potential—connecting events. Thus, **c** occurs at time t, say, and initiates two processes, one of which results in the occurrence of **d**, and the other of which would have resulted in the occurrence of **b**—and thereby prevented the process proceeding from **d** from resulting in **e**—had it not been interrupted by the process initiated by **a**. So **a** and **c** are the *only* events occurring at t; of these, **c** is clearly the only producer of **e**. But again, the unique minimally sufficient set for **e** is {**a**, **c**}.

We cannot get **c** to come out as a cause of **e** by appealing to the Intrinsicness thesis, searching for a sequence of events just like the **c**-**d**-**e** process that will serve as a "blueprint". For the laws are such that if any event just like **c** occurs, it initiates two processes, one of which (the **b**-process) will cut short the other if not itself interrupted. That means, in effect, that the **c**-**b**-**e** connections in figure 13 cannot be "subtracted" on pain of violating the laws (as they could, if figure 13 depicted neurons physically connected by stimulatory and inhibitory channels). The only obvious alternative is to invoke the Transitivity thesis—for it is easy enough to get **c** to come out as a producer of **d** ({**c**} is the unique minimally sufficient set for **d**), and **d** of **e** ({**d**} is likewise the unique minimally sufficient set for **e**). And that means that Transitivity had better be built directly into the analysis.

As to Intrinsicness, consider again the tale of Suzy, Billy, and the broken bottle. Let **c** be one of the events constituting the trajectory of Suzy's rock. Let **a** be an event in the trajectory of Billy's rock that is contemporaneous with **c**. And to keep things clean, let us suppose that Billy has thrown a Smart Rock, programmed to guarantee that the bottle shatters at just the time and in just the manner it does. That means that {**a**} is sufficient (and therefore minimally sufficient) for **e**, no matter how modally precise we take **e** to be.[22]

And that means that **c** does not belong to a unique minimally sufficient set for **e**. So **c** does not yet come out as a producer of **e**. Furthermore, invoking Transitivity is no help, since the argument works no matter how causally proximate **c** and **a** are to **e**. The only obvious alternative is to invoke Intrinsicness, using as a blueprint the sequence of events that would have occurred if Billy had not thrown his rock. And that means that Intrinsicness had better be built directly into the analysis as well.

Let us next consider the causal status of omissions, in light of our provisional analysis of production.

First, it is a direct consequence of the analysis that no omission can help produce an event (be that event an omission itself or not). For if any omission is a producer, then there must be at least one example of an event **e** and a set of contemporaneous events S (all occurring at earlier time t, say) such that S is minimally sufficient for **e**, and S contains at least one omission **o**. But there are no such examples.

Proof: Suppose S and **e** are such an example. Since S is minimally sufficient for **e**, it is sufficient for **e**; so it must follow from the laws, together with the premise that all the events in S occur at t, together with the premise that no other events occur at t, that **e** occurs. And recall that the "no other events occur" clause must be understood as quantifying over *genuine* events, not omissions. But in that case, the set S - {**o**} must *also* be sufficient for **e**; for the claim that the events in this set occur, but no other *genuine* events do, is *equivalent* to the claim that the events in S occur (including **o**), but no other genuine events do. The stipulated occurrence of **o**, in the latter claim, is, as it were, simply "swallowed up" in the part of the former claim that states that no other genuine events occur. It therefore follows that S is not, contra hypothesis, minimally sufficient for **e**.

With modest assumptions about the laws, we can also prove that no omission can be produced—that is, there is no omission **o** and event **c** such that **c** helps to produce **o**. For in order for some omission to be produced, there must be at least one example of an omission **o** and a set of events S such that S is minimally sufficient for **o**. Now for the assumption: The laws of evolution are such that the unique state of the world in which *nothing at all* happens—no event occurs—remains unchanged, evolving always into itself. If the laws are like this, then S cannot possibly be minimally sufficient for **o**, simply because a proper subset of S—namely, the empty set—will be sufficient for **o**.

Let us finally consider more closely why production and dependence so often coincide. First, suppose that **c**, which occurs at t, is a producer of **e**. In *typical* cases—that is, if the environment does not conspire in such a way as to ruin the ordinarily neat nomological relationships between **c** and **e**—**c** will belong to a unique minimally sufficient set for **e**. Let S be this set. Other events occur at t; let us collect these together into the set T. Then consider what happens in a counterfactual situation where **c** does not occur (keeping in mind that there may be more than one such situation): (i) The other events in S occur. (ii) The events in T occur. (iii) Possibly, in place of **c**, some other event **c**' occurs.[23]

Let S' = S – {**c**}. Then—modulo a small assumption—S' ∪ T cannot be sufficient for **e**. For suppose it is. Then given (the small assumption) that not every subset of S' ∪ T that is sufficient for **e** contains a proper subset that is also sufficient for **e** (given, that is, that the sufficient subsets of S' ∪ T are not infinitely nested), S' ∪ T will contain a minimally sufficient set for **e**. But then S will not be the unique such minimally sufficient set, contra hypothesis.

So it will not follow from the premise that all the events in S' ∪ T occur, together with the premise that no other events occur, together with the laws, that **e** occurs. And that means that in a counterfactual situation in which **c** does not occur, *and in which no event takes its place*, **e** does not occur. That is one way it could turn out that **e** counterfactually depends on **c**.

More likely, though, some event **c**' *will* take the place of **c**. Furthermore, if this event conspires in the right way with the other events in S' ∪ T, then **e** will occur all the same. Suppose for example that in the actual situation, Suzy throws a rock, breaking a window. Billy is absent this time, so we can assume that her throw is part of a unique minimally sufficient set for the breaking. But suppose further that if she hadn't thrown, her hand would (or simply *might*) have fallen by her side, brushing against a switch, flipping it and thereby activating a catapult that would have hurled a brick at the window, breaking it. Then that is a situation in which the counterfactual alternative to **c** (or: one of the alternatives) conspires with other events to bring about **e**. But notice that it takes some work to rig an example so that it has this feature; *typically*, when **c** is part of a unique minimally sufficient set for **e**, it will be the case that if **c** hadn't occurred, then whatever event replaced it would *not* so conspire. Which is to say that it will be the case that **e** depends on **c**.

Finally, even in the case of Suzy, the rock, and the stand-by catapult, we will have *step-wise* dependence of the window's breaking on her throw: Picking an event **d** that forms part of the rock's flight, we will have dependence of the breaking on **d** and of **d** on the throw. We could tinker further to destroy one of these dependencies, but of course that is not enough: We would need to tinker enough to block *every* two-step chain of dependencies—and every three-step chain, and every four-step chain, … etc. It is possible to do this—even while guaranteeing the existence, at each stage, of a unique minimally sufficient set for the window's breaking—but only at the cost of making the example even more atypical.[24] And that shows that if I am right about the correct analysis of the central kind of causation, then it is no great surprise that the simple idea that causation should be understood as the ancestral of counterfactual dependence worked as well as it did.

§8 Applications, open questions, and unfinished business

§8.1 Applications

In the last three decades or so, causation seems to have become something of a philosophical workhorse: philosophers have offered causal accounts of knowledge, perception, mental content, action, explanation, persistence through time, and decision making, to name a few. It won't be possible to discuss any of these topics in detail, but I will focus briefly on the latter three as a way of beginning to explore the broader consequences of the distinction between dependence and production.

Before doing so, however, let us just observe that even the most cursory inspection of the philosophical roles causation plays vindicates one of the three central arguments for the distinction, which is that Transitivity and Dependence conflict. Recall one of the examples used to display the conflict: Billy spies Suzy, and runs towards her in an effort to stop her from throwing a rock at a window; en route he trips, and as a consequence doesn't reach her in time to stop her. The window breaks. If Billy hadn't tripped, the window would not have broken (because he would have stopped Suzy). If Billy hadn't run toward Suzy, he wouldn't have tripped. Suppose we conclude, via a confused appeal to Dependence and Transitivity, that Billy's running toward Suzy was one of the causes of the window's breaking. Still, we will have to admit that for a 'cause' that is so proximate to its 'effect', it is quite strange: It is not something we would cite as part of an explanation of the breaking, we would not hold Billy at all responsible for the breaking on account of having helped to 'cause' it in this way, etc. In short, consider any of the typical roles that causation plays in other arenas, and you will find that the sort of relation Billy's action bears to the breaking quite obviously plays none of those roles. That should add to the conviction (if such addition were needed) that this relation is not one of causation.

That is not to say that we cannot describe this relation in causal terms, since of course we can: Billy does something which both (i) initiates a process that threatens to interrupt the window-breaking process, and (ii) causes an event that interrupts this potential interrupter. So the relation of Billy's action to the window-breaking has a perfectly definite causal structure. But that does not make it a kind of causation.

If we look more closely at some "causally infused" concepts, I think we can find more direct manifestation of the difference between production and dependence, even in the kind of brief and selective treatment I am about to offer.

Begin with persistence. On one well-known view, what it is for an object to persist from time t_1 to time t_2 is for it to have temporal parts at t_1, t_2, and the intervening times such that earlier ones are appropriately connected to later ones. What I want to focus on is not the controversial ontology of temporal parts, but the nature of the connection—which, on typical formulations, has got to be partly *causal*. The question is: Could the causal relation involved in this connection be one of mere dependence, without production?

Good test cases are not easy to come by, mainly because we already know that for an enduring object of any complexity, the causal component of the connection between its earlier and later stages has to be understood as much more restrictive than *either* dependence or production, and the restrictions are not easy to spell out.[25]

Let's strive for as simple a case as possible—say, one involving the persistence through time of an electron (assuming for the moment a naïve conception of the electron as a classical point-particle). Suppose we have two electron-stages, located at t_1 and t_2 ($t_1 < t_2$). Plausibly, it is a necessary condition for the stages to be stages of the same persisting electron that the first be a cause of the second. But I think this necessary condition cannot possibly be met if the second merely *depends* on the first. Suppose, for example, that the presence of the first results in the prevention of something which would itself have prevented the presence of the second. If we know that that is the *only* causal connection that obtains between the two stages, then we know enough to conclude that they cannot be stages of one and the same electron. On the other hand, if we know that the first stage helps to produce the second stage, then while we may not yet know enough to conclude that they belong to the same electron, it does *not* matter whether we learn that the second stage fails to depend on the first (as it might, because of some backup process that would have led to an electron in the same place at the same time).

Thus, while the matter certainly merits further investigation, we can conclude that with respect to persistence through time of a simple object like an electron, production is the important causal notion (or at least: one of the basic ingredients in this causal notion), whereas dependence is irrelevant. Persistence of more complex objects seems unlikely to differ in this respect.

Consider next an arena in which the relative importance of the two kinds of causation is reversed: causal decision theory. When you face a range of options, and causal decision theory says (very roughly) that the rationally preferable one is the one most likely to have as a causal consequence the best (by your lights) outcome, the notion of "causal consequence" at work is clearly that of dependence, and not

production. Or rather, it is a natural generalization of dependence, where we allow that more than just *events* can be suitable relata: facts, say, or states of affairs.

Our standard stories of double-prevention already illustrate the irrelevance of production to decision-making. There is, we can suppose, nothing whatsoever that Billy can do to help *produce* the bombing, but that doesn't matter in the slightest: For whether there is a bombing clearly *depends* on the action he takes, and it is his beliefs about this dependence and its detailed structure that will guide his decisions, in so far as he is rational.

Here is another kind of example that makes vivid the irrelevance of mere production; once you see the trick you can generate endless variations: You want your team to win; that is the only thing in the world that matters to you. You know that if you do nothing—just stay where you are, sitting on the sidelines—your fellow teammates will, with certainty, achieve victory. On the other hand, if you insist on playing, the team will probably lose (you are not, alas, very good). Still, if you play and the team manages to *win*, then your own actions will have helped to produce that win (you have your good days, and with luck this might be one of them). If what matters in decision is what your different courses of action would be likely to *produce*, then you should play—for only then does your action have a chance of helping to produce something desirable. That you should clearly *not* play helps show that the productive upshot of your actions is not what matters. (Don't say: "But if you sit on the sidelines, then you help produce the victory by not playing." That is to confuse the kind of causation that omissions can enter into with production, and we have already seen ample reason why such confusion should be avoided.)

There is a needed qualification that by now might be obvious. For there is an obvious way that production *can* matter, quite a lot, to decision: namely, if the outcomes to which the agent attaches value are themselves partly characterized in terms of what produces what. To modify our example, suppose that what matters to you is not that your team win, per se, but that you *help bring about* your team's victory. If so, then it will be perfectly rational (though selfish) for you to insist on playing. Again, if something awful is going to happen no matter what you do, it may yet matter quite a lot to you whether it happens partly as the productive upshot of your behavior.

So production can matter to decision, after all. But seen as an objection to my thesis this is just an equivocation, since that thesis concerns the kind of causal relation that connects action to outcome, and not the taxonomy of outcomes themselves. Even when you choose to avoid a certain course of action because it would result in *your* having helped produce the evil deed, the sense of "because" is clearly that of dependence: What matters is how the possible outcomes (evil deed which you helped produce vs. evil deed in which you had no hand) *depend* on your action.

Let us finally consider an arena—causal explanation—in which *both* production and dependence play a role; what we'll find is that these roles are interestingly different.

Recall, once more, the story of Billy's failed attempt to stop Suzy from breaking the window. In seeking an explanation of the window's shattering, we might, on the one hand, ask what *brought that event about*, what *led up to it*. To questions of this sort, it would be strange, to put it mildly, to cite Billy's trip; rather, what's wanted is information about the *producers* of the shattering. On the other hand, we can ask why the shattering occurred, *given that* Billy started running toward Suzy. When we ask a question of this form—"Why did **e** occur, given that **c** occurred"?—we are obviously presupposing that **c** set in motion some process or processes that would have prevented **e**, had circumstances been different—and so we want to know about events whose occurrence *kept* circumstances from being different. (Sometimes we make the presupposition quite explicit: "Why did the window break, given that Billy set out to stop Suzy?") With respect to our story, the obviously correct answer is that Billy tripped (or that he wasn't watching where he was going, or that he's clumsy and so prone to tripping, etc.). It *won't* do to cite an arbitrary producer or **e**, unless it also happens to play the role of stopping **c**'s occurrence from preventing **e**'s. Try it: "Why did the window shatter, given that Billy started running toward Suzy (with the intention of stopping her)?" "Because Suzy threw a rock at the window." Highly misleading, to say the least.

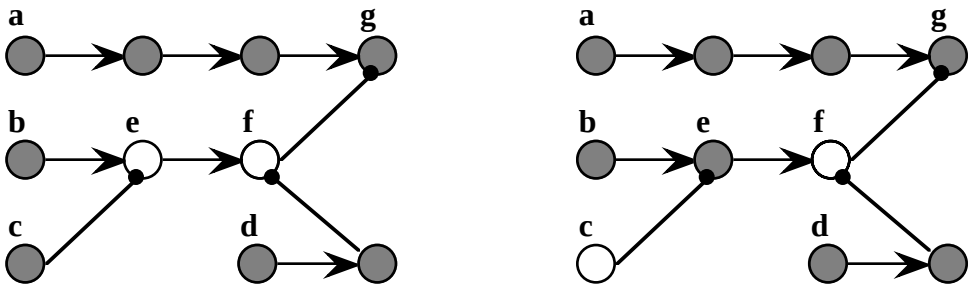
The catch-all explanation-request that philosophers often focus upon—"Why did the window break?"—obscures the difference between these two more refined ways of requesting an explanation. Indeed, in the right contexts, the question "Why did the window break?" might be appropriately answered either by "Billy tripped" or by "Suzy threw". That shows that both dependence and production have causal-explanatory roles to play. But it doesn't show—indeed it hides the fact—that these roles exhibit interesting and striking differences. There is, after all, a world of difference between asking, of some event, what led up to it, and asking why it occurred, given that something else was poised to prevent it—never mind that each question could, in the right context, be conveyed by "Why did it happen?"

That concludes my necessarily brief discussion of the implications that the distinction between production and dependence might have for other areas of philosophy. I hope it has been detailed enough

to make further such inquiry seem worthwhile. I will close now with a brief look at a few ways in which the picture of causation that emerges from the foregoing treatment turns out to be more complicated than one might have thought.

§8.2Unfinished business

First, there are certain kinds of cases that we have some inclination to call cases of causation, but that also elude classification in terms of production or dependence. Here is an example, a slight variation on the story of Billy, Suzy, and Enemy: This time, there is a second fighter plane escorting Suzy. Billy shoots down Enemy exactly as before, but if he hadn't, the second escort would have. Figure 14 captures the salient causal structure:



c fires if **c** hadn't fired…

Figure 14

It is no longer true that if Billy hadn't pulled the trigger, then the bombing wouldn't have happened. In figure 14, it is no longer true that if **c** hadn't fired, then **g** wouldn't have. Nevertheless, given that Billy's action is partly responsible for the success of the bombing in the first case, where the second escort was absent, then surely there is some inclination to grant him such responsibility in this second case, which merely adds an alternative that plays no active role. In the diagram, the superfluous preventive chain from **d** should not, it seems, change **c**'s status as a kind of cause of **g**.

Notice that our judgment that Billy is partly responsible for the success of the bombing is quite sensitive to the nature of the backup preventer. For example, suppose that Suzy is protected not by a second escort, but by a Shield of Invulnerability that encloses her bomber, making it impervious to all attacks. We have the same relations of counterfactual dependence: if Billy had not fired, the bombing would still have been a success, but if Billy had not fired and the Shield had not been present, the bombing would not have been a success, etc.. But only with great strain can we get ourselves to say that, in this case, Billy is partly responsible for the success of the bombing.

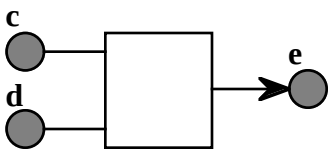
The issue here is really exactly the same as one that has received some discussion in literature, concerning the nature of prevention and how best to analyze it. Thus, McDermott (1995) asks us to consider a case in which a catcher catches a ball flying towards a window; the window, however, is protected by high, thick brick wall, and so of course would not have broken even if the catcher had missed her catch. Does the catcher prevent the window from being broken? It seems not. But if we replace the wall with a second catcher—one who would have caught the ball if the first catcher had missed—then judgments tend to be reversed. All I have done with the case of Billy, Suzy, and the backup escort is to insert such a case of "preempted prevention" into the middle of the story. In short, the tricky problem of how to understand the exact nature of preempted prevention generates, as a kind of side-effect, a problem for how to understand certain kinds of double prevention (namely, where the first preventer has a preempted back up). (For a sensitive and insightful treatment of the problem of preempted prevention, see Collins 2000.)

Here is one possible explanation for what is going on in these cases[26]: When we judge that Billy is partly responsible for the success of the bombing, that is because we are treating the bombing as counterfactually dependent on his action—not, admittedly, on the ordinary way of understanding the counterfactual, but on a slightly different way that holds slightly different facts about the scenario fixed. That is, in the actual scenario the back-up escort does not, let us suppose, fire on Enemy. If we hold *that* fact fixed when evaluating the counterfactual, then we get the result that if Billy had not fired, then (holding fixed the fact that the back-up escort does not fire) the bombing would not have succeeded. Moreover, this reading of the counterfactual seems permissible (although not obligatory). On the other hand, a parallel reading of the counterfactual in the version of the story where Suzy is protected by a Shield of Invulnerability seems strained: for what extra fact is to be held fixed? Presumably, we have to

hold fixed the fact that the Shield does not repel Enemy's missiles—but it is not at all clear how to construct (without overly gratuitous deviations from actuality) a counterfactual situation in which Billy does not fire, but this is the case.

I am not at all sure about the prospects for this proposal, and hence take it that solving the problem of preempted prevention is a piece of unfinished business that affects my account of causation, by way of complicating the "dependence" half of it.

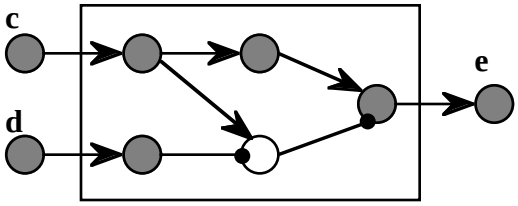
A different issue arises from cases that look like production until they are examined up close. I will consider just one such case, and then tentatively offer a lesson that I think we should draw from it. The boxed neuron in the diagram below functions as an "AND" gate: When and only when "input" neurons **c** and **d** both fire, it fires, causing **e** to fire:



AND gate

Figure 15

A paradigm case of the production of one event by two other events, it would seem. But is it? Note that I did not tell you what it was for the boxed neuron to fire, what such firing consists in. Here's a closer look at the inner workings of the box:



AND gate, close-up

Figure 16

Now it appears that **d** is not a producer after all. We could leave the matter there, with the fairly obvious observation that our judgments about the causal structure of this set-up are, of course, going to be defeasible, given further information about the detailed workings of the set-up. But I think a different and somewhat more interesting lesson is called for. Notice first that it would be inapt to say that **d** is not a producer of the *firing of the box*; it is rather that *within* the chain of events that constitutes the firing of the box, the incoming signal from **d** plays the role of double-preventer. It does not, for example, help produce the firing of the right-most mini-neuron inside the box. Now, that observation might seem to be of little import. For example, it doesn't seem to bear at all on the question of what the relationship between the firing of **d** and the firing of **e** is. Is that production or dependence?

But in fact I think the observation is relevant. For I think it is correct to say that when **c** and **d** both fire, the firing of **d** helps produce the firing of the "AND" neuron, but I also think it is correct to say that the firing of **d** does not help to produce the firing of the right-most constituent neuron in the box. So far that leaves us some room for maneuver with respect to the firing of **e**: We could, for example, adopt the view that if **d** helps produce *any* event that is itself a producer of **e**, then by transitivity **d** helps produce the firing of **e**. Then **d** comes out as a producer of **e**, since it helps produce the firing of the "AND" neuron, which in turns helps produce the firing of **e**. Alternatively, we could adopt a much more stringent standard, and say roughly that in order for **d** to be a producer of **e**, *every* event causally intermediate between the two must be produced by **d** and producer of **e**. Since the firing of the right-most neuron in the box is such a causally intermediate event—and since **d** does not help produce it—**d** will not come out as a producer of **e**.

What I think we should say is that depending on context, either answer can be correct. More specifically, I think that causal judgments are tacitly relative to the level description we adopt when giving an account of the relevant chain of events (and that this choice of level of description will be a feature of the context in which we are making our causal judgments). In giving an account of the events in

figure 15, for example, we can adopt a level description that includes such categories as "firing of the 'AND' neuron"; that is, we will provide some such description as "neurons **c** and **d** both fire, each sending a stimulatory signal to the boxed neuron, which then fires, emitting a stimulatory signal which reaches **e**, which then fires". Alternatively we can adopt a level description which speaks not of the firing of the "AND" neuron, but rather of the various firings of the neurons within it and of the stimulatory and inhibitory connections between them. Relative to the first choice of level of description, **d** comes out as a producer of **e**; relative to the second, it does not.[27]

Obviously, it is another major piece of unfinished business to spell out the relevant notion of "levels of description", and to explain exactly how such levels find their way into the contexts in which we make our causal judgments. I'll leave that business unfinished, and content myself with responding to one objection. That is, I realize that some will view this introduction of context-sensitivity into the account as a serious cost—but as far as I can see such an attitude manifests the same confusion we saw earlier of a sound methodological precept for an unsound *a priori* conviction about the workings of our causal concepts. Just as it makes good methodological sense to begin an investigation of our concept of causation with the working hypothesis that it is univocal, so too it makes good sense to adopt as an initial working hypothesis the view that it is not context-relative. But there's no sense whatsoever in maintaining such hypotheses when our investigations have revealed complexities too serious for them to accommodate. If I am right, the view we are pushed to is that our thinking about causation recognizes two basic and fundamentally different varieties of causal relation, and that which relation is in play in any given situation is—or least can—depend on contextually-specified features of how we are conceptualizing that situation. That this is a view that quite obviously needs detailed argument and defense should make it seem unattractive only to those with an excessive devotion to the curious notion that philosophical life should be easy.

REFERENCES

Bennett, Jonathan 1987: "Event Causation: The Counterfactual Analysis," reprinted in Sosa and Tooley, ed., *Causation*, pp. 217-33.

Bennett, Jonathan 1988: *Events and Their Names.* Hackett.

Collins, John 2000: "Preemptive Prevention". *Journal of Philosophy,* 97, pp. 223-34.

Hall, Ned 2000a: "Causation and the Price of Transitivity". *Journal of Philosophy,* 97, pp. 198-222.

Hall, Ned 2000b: "Non-locality on the cheap?". ms.

Hall, Ned 2000c: "The intrinsic character of causation". ms.

Kim, Jaegwon 1973: "Causes and Counterfactuals". *Journal of Philosophy,* 70, pp. 570-2.

Lewis, David 1973a: *Counterfactuals*, Blackwell.

Lewis, David 1973b: "Counterfactuals and Comparative Possibility," *Journal of Philosophical Logic* 2, pp. 418-46.

Lewis, David 1979: "Counterfactual Dependence and Time's Arrow," *Noûs* 13, pp. 455-76; reprinted with "Postscripts" in his *Philosophical Papers,* Vol. II, pp. 32-66 (page references are to this latter printing).

Lewis, David 1986a: "Causation". Reprinted with postscripts in his *Philosophical Papers, Vol. II*, Oxford: Oxford University Press, pp. 159-213.

Lewis, David 1986b: "Events". In his *Philosophical Papers, Vol. II*, Oxford: Oxford University Press, pp. 241-69.

Lewis, David 2000: "Causation as Influence". *Journal of Philosophy,* 97, pp. 182-97.

Lombard, Lawrence 1990: "Causes and Enablers". *Philosophical Studies* 59.

Mackie, J. L. 1965: "Causes and Conditions". *American Philosophical Quarterly,* 2/4, pp. 245-64.

Maudlin, Tim 2000: "A Modest Proposal Concerning Laws, Counterfactuals and Explanations". ms.

McDermott, Michael 1995: "Redundant Causation". *British Journal for the Philosophy of Science,* XLVI, pp. 523-44.

Mellor, D. H. 1997: *The Facts of Causation.*

Paul, L. A. 1998: "Keeping Track of the Time: Emending the Counterfactual Analysis of Causation". *Analysis* 58(3), pp. 191-198.

Schaffer, Jonathan 2000: "Trumping Preemption". *Journal of Philosophy,* 97, pp. 165-81.

Sosa, E. and M. Tooley eds. 1993: *Causation.* Oxford: Oxford University Press.

Stalnaker, Robert 1968: "A Theory of Conditionals," in Nicholas Rescher, ed., *Studies in Logical Theory*, Blackwell.

Swain, Marshall 1978: "A Counterfactual Analysis of Event Causation". *Philosophical Studies* 34: 1-19.

---

[1] See for example Mellor 1997 and Bennett 1988.

[2] It is hardly a peculiarity of the counterfactual analysis that it needs this sort of supplementing. Suppose, to cite a familiar story (Adapted from Mackie 1965), that we analyze a cause of an event as a distinct event which is sufficient, given the laws and relevant circumstances, for that first event's occurrence. This analysis can be "refuted" in an equally trivial fashion by noting the following "consequences": the slamming causes the shutting; the shattering's constituent events jointly cause it (and, perhaps, it causes them as well); there is still a fairly immediate common cause of the two widely separated shatterings (namely, the "conjunctive" event which, necessarily, occurs iff both Suzy's and Billy's throws occur); the "emailed two days ago" event **c** still causes the earlier reply **e**. So I do not think that these (admittedly challenging) issues involving the nature and individuation of events pose any problem for counterfactual analyses of causation *in particular*.

[3] For excellent discussions of the issues involved in providing a full-blown philosophical account of events, see Lewis (1986b) and Bennett (1988).

[4]Notice also that Bennett and Lewis both use the locution "A causes B" rather than the weaker "A is *a* cause of B"—thus illegitimately suggesting that a particularly *salient* causal connection is being asserted. That won't do; after all, doesn't it also sound wrong to say, e.g., that the forest's presence caused the fire? But it doesn't sound so bad to say that it was *a* cause of, or *among* the causes of, the fire.

[5]Notice that the distracting intuitions evoked by Bennett's example are silent here: There is no "good bit of common sense" analogous to Lombard's observation that "heavy rains … don't start [fires]"; furthermore, no event stands out as a particularly salient cause of the forest's presence (although in the right context, the April rains just might!).

[6]See his 1986a.

[7]For standard treatments of the counterfactual, see, e.g., Stalnaker 1968, Lewis 1973a, and Lewis 1973b. Whether these standard treatments are adequate to the needs of the counterfactual analysis is a question we will take up shortly.

[8]More precisely, where X and Y are propositions, the conditional $X \rightarrow Y$ is a backtracker iff the following hold: X is entirely about a time or times later than the time or times which Y concerns; and Y is in fact false. Intuitively,

such a conditional says that if things at one time had been a certain way, then things at an earlier time would have been different from how they actually are. Note that my definition does not quite coincide with Lewis's original definition; see his 1979. The differences, however, are unimportant for our present purposes.

[9] Note that Lewis's account of the counterfactual conditional does not rule out (what I have called) backtrackers in principle, but only when the world exhibits an appropriate sort of (contingent) global asymmetry. He thus leaves room for the possibility of backwards causation.

[10]Swain (*op. cit.*; see especially pp. 13-14) has overlooked this point. He considers a case with exactly the structure of that described by figure 2, yet fails to notice that his views on counterfactuals deny him the resources needed to secure the link between **c** and **e**.

[11]In saying this, I am taking sides in a debate over the correct analysis of counterfactuals. Thus Stalnaker argues that the denial of $X \rightarrow Y$ is equivalent to the contrary conditional $X \rightarrow \sim Y$. But if he is right, then my job is easier: for I aim here merely to show that, in the case described by figure 2, the defender of the counterfactual analysis must assert the conditional $\sim O\mathbf{d} \rightarrow O\mathbf{c}$. If Stalnaker is correct in claiming that he does this automatically when he denies the backtracker $\sim O\mathbf{d} \rightarrow \sim O\mathbf{c}$, then so much the better for me!

[12] See for example Lewis, 1979.

[13] See for example Maudlin 2000.

[14] The usual caveats apply: There might be more than one such minimal alteration to the state at t, or there might be an infinite sequence of minimal alterations, each more minimal than its predecessor. Either way, the proper fix is to take the conditional to be true just in case there is some alteration A such that the consequent comes out true for every choice of alteration A' which is at least as minimal as A. Note also that we would need to amend this rule, if we wanted our analysis to accommodate backwards causation and action at a temporal distance.

[15] But see my 2000c for detailed argument and discussion.

[16] The word "structure" is intentionally ambiguous: We could take the structure to be the mereological fusion of all the events, or we could take it to be a set-theoretic construction out of them (most simply, just the set of them). It doesn't matter, as long as we're clear on what duplication of event-structures amounts to.

---

[17] For various reasons, not worth the long digression their spelling out would require, I do not think we can add an "only if" to the "if" to get "iff". See my "Non-locality on the cheap?" for discussion.

[18] There are, in the literature, various other apparent counterexamples to Transitivity that cannot be handled merely by denying Dependence (see for example McDermott 1995). But on my view they are only apparent; see my 2000a for detailed discussion.

[19] I do not think that *whenever* we have counterfactual dependence between two facts, the statement asserting this dependence should be construed as causal. It depends on why the dependence holds. For example, "If it hadn't been that P, it wouldn't have been that Q" could be true because Q entails P, in which case we shouldn't view this sentence as expressing a *causal* truth. Alas, I don't think we can hope to circumscribe the *causal* dependence claims merely by demanding that the facts in question be logically independent; for what of counter-nomics, such as "If gravity had obeyed an inverse-cube law, then the motion of the planets would not have obeyed Kepler's Laws?" While it seems intuitively clear when we have a relation of dependence that holds for the right sorts of reasons to count as causal, and when we don't, I will leave the project of elucidating these reasons for another occasion.

[20] Observe, however, that (iii) does not follow from (iv): It is perfectly consistent, for example, to claim that there is but one kiss, but that it could have been more perfunctory than it was. Also, (iii) should not be confused with the overly strong claim that *all* of an event's properties are essential to it. On any reasonable view, Suzy's kiss—even if essentially passionate—is only accidentally such as to upset her mother.

[21] I am also assuming that the physical state of the world has no emergent features. Alas, quantum mechanics is probably an exception, given that the physical states of composite systems do not, in that theory, supervene on the physical states of their component parts. I won't pursue here the complications that this issue raises.

[22] Strictly speaking, we should include various other events along with **a**: for example, the events contemporaneous with **a** that constitute the bottle's presence, then. Including them won't make any difference to the argument.

[23] Consider, for example, Suzy's throw, in the simple situation where Billy's doesn't also throw: had this throw not occurred, it is not that *nothing* would have happened in its place; rather, her arm would have (say) simply remained at her side. That's an event, even if not one ordinarily so called.

---

[24] For example, imagine that there is some exquisitely sensitive alarm system, that will be triggered if the rock deviates at all from its actual path; triggering it in turn initiates some other process that will break the window. It is possible to set up such an example so that at each moment t of the rock's flight, there is a unique minimally sufficient set for the breaking (a set which will include the part of the flight that is occurring at t); but plausibly the breaking will depend on not one of the events making up the flight.

It's worth contrasting this sort of failure of dependence with the more usual cases of preemption: In those cases, there is in addition to the process that brings about the effect some other, rival process that somehow gets cut off. Here, by contrast, there *is* no such rival process; there only *would* be, if the events in the main process were different. We might therefore call these cases of overdetermination by a "preempted" merely *potential* alternative. They pose special problems for counterfactual analyses, since many of the strategies for dealing with cases of overdetermination by preempted *actual* alternatives fail to apply.

[25] You step into the machine, and as a consequence (i) your body dissolves; (ii) the machine transmits a signal to a second, distant machine. It receives the signal, and as a consequence produces a body that is an exact replica of yours, as it was when you stepped into the first machine. Is it you? Even those of us who believe that teletransportation is possible can't say "yes", without knowing more of the causal details. For example, it might be that the second machine had been ready for a long time to produce a body—one that just by chance happened to be exactly like yours—and that the signal from the first machine merely acted as a catalyst. Then the two person-stages are stages of different people, even though the first is a cause, in both senses, of the second.

[26] Inspired by a proposal of Steve Yablo's (personal communication).

[27] Although I will not pursue the matter here, it is worth noting that it is quite easy to accommodate this relativity to level description within the analysis of production offered in §7.