Counterfactuals and causation: history, problems, and prospects

(Chapter 1 in Collins, Hall, and Paul eds, *Counterfactuals and Causation*)

John Collins, Ned Hall, and L. A. Paul

Among the many philosophers who hold that causal facts[1] are to be explained in terms of—or more ambitiously, shown to reduce to—facts about *what happens*, together with facts about the *fundamental laws* that govern what happens, the clear favorite is an approach that sees *counterfactual dependence* as the key to such explanation or reduction. The paradigm examples of causation, so advocates of this approach tell us, are examples in which events *c* and *e*—the cause and its effect—both occur, but: had *c* not occurred, *e* would not have occurred either. From this starting point ideas proliferate in a vast profusion. But the remarkable disparity among these ideas should not obscure their common foundation. Neither should the diversity of opinion about the prospects for a philosophical analysis of causation obscure their importance. For even those philosophers who see these prospects as dim—perhaps because they suffer post-Quinean queasiness at the thought of *any* analysis of *any* concept of interest—can often be heard to say such things as that causal relations among events are somehow "a matter of" the patterns of counterfactual dependence to be found in them.

It was not always so. Thirty-odd years ago, so-called "regularity" analyses (so-called, presumably, because they traced back to Hume's well-known analysis of causation as constant conjunction) ruled the day, with Mackie's *Cement of the Universe* embodying a classic statement. But they fell on hard times, both because of internal problems—which we will review in due course—and because dramatic improvements in philosophical understanding of counterfactuals made possible the emergence of a serious and potent rival: a counterfactual analysis of causation resting on foundations firm enough to be repel the kind of philosophical suspicion that had formerly warranted dismissal. One speculates at peril about the specific intellectual and social

---

[1] Let us be clear from the outset that by "facts" here we mean nonlinguistic items.

forces driving trends in philosophy. Still, it is a safe bet that Lewis's groundbreaking paper "Causation" (1973a) helped to turn the tide. Elegant, concentrated, and compelling, this paper contained a forceful condemnation—much quoted, by this point—of the incumbent view:

> It remains to be seen whether any regularity analysis can succeed in distinguishing genuine causes from effects, epiphenomena, and preempted potential causes—and whether it can succeed without falling victim to worse problems, without piling on the epicycles, and without departing from the fundamental idea that causation is instantiation of regularities. I have no proof that regularity analyses are beyond repair, nor any space to review the repairs that have been tried. Suffice it to say that the prospects look dark. I think it is time to give up and try something else.
>
> A promising alternative is not far to seek. (1973a, p. 160)

At the time, the sentiment Lewis expressed in this passage was entirely fitting.

Whether—and to what extent—it continues to be fitting is a judgment that this volume aims to inform. Here we bring together some of the most important recent work connecting—or in some cases, disputing the connection between—counterfactuals and causation. Some of these essays (chapters 2 - 4, 7, and 8) appeared previously in a special issue (April 2000) of the *Journal of Philosophy* devoted to this topic; the rest are published here for the first time, or in new and updated forms. (Lewis's "Causation as Influence"—chapter 3—has been greatly expanded for this volume.) Our aim in this introductory chapter is to set the stage for the papers that follow, by providing a map of the main features of the philosophical terrain that our authors are navigating. Accordingly, we will discuss the work on the semantics of counterfactuals that set the stage for counterfactual analyses of causation (section 1), review the early successes and more recent challenges to such analyses (sections 2 - 4), probe some important points of methodology that play a significant but often tacit role in philosophical investigations of causation (section 5), and review (in sections 6 and 7) two questions of quite general interest that work in the counterfactual tradition has shed a great deal of light on, and that many of our contributors are concerned to address: whether causation is transitive, and what the nature of its relata is.

For those of you who come to our topics for the first time, our overview should provide you with the background you need to appreciate the often subtle and intricate dialectical interplay featured in the remaining contributions; for those of you who come as seasoned veterans, we hope

nevertheless to reveal unnoticed connections within and novel perspectives on what may have seemed boringly familiar material. (Those of you who are impatient can skip ahead to section 8, where we give capsule summaries of the contributions in this volume.)

Let us begin by rehearsing the first half of our topic: counterfactuals.

1. Counterfactuals

A *counterfactual* is a conditional sentence in the subjunctive mood. The term "counterfactual" or "contrary-to-fact" conditional carries the suggestion that the antecedent of such a conditional is false. Consider for example: "If this glass had been struck, then it would have shattered". The implication is that the glass was actually not struck and did not shatter. Yet these conditionals do not appear to differ in kind from those in which the truth-value of the antecedent is an open question (e.g. "If this glass were struck, then it would shatter"). While this suggests that the expression "subjunctive conditional" might more appropriately delineate the topic, the term "counterfactual" is by now so well-entrenched that it will suffice to stipulate that counterfactuals may have true antecedents.

The philosophical significance of counterfactuals is apparent from the way in which they have figured so prominently in recent discussions of many other philosophically important concepts: knowledge, perception, and freedom of the will to name just a few. An attractive thought is that this conceptual connection is best understood as being mediated by the concept of *cause*. Thus, for example, counterfactuals are relevant to discussions of perception because the most promising theories of perception are causal theories, and counterfactuals are fundamental to any philosophical understanding of causation.

The thought from which counterfactual theories of causation proceeds is this. A certain glass is struck, and shatters. To say that the striking of the glass caused the shattering of the glass is to say that if the glass had not been struck then it would not have shattered. The striking caused the shattering in virtue of the fact that the shattering was *counterfactually dependent* on the striking. The papers in the present volume are devoted to exploring the prospects of an account of causation

based on that simple idea. But before turning to that discussion, we shall review some of the basic semantic features of counterfactuals.

For counterfactual conditionals themselves pose a problem that must to be solved before such conditionals are recognized as a fit tool for further philosophical analysis. The problem, first raised in the recent philosophical literature by Chisholm (1946) and Goodman (1947), is to provide non-circular truth conditions for the general counterfactual conditional with antecedent A and consequent C. We shall abbreviate this conditional as "A >> C", to be read: "If it were the case that A, then it would be the case that C".

Observe first of all that the counterfactual must be distinguished from both the material conditional and strict entailment. The material conditional has truth conditions that are too weak; not every subjunctive conditional with a false antecedent is true. But on the other hand the strict conditional is too strong; the shattering of the glass is not logically entailed by the striking of the glass.

The right account is, presumably, located somewhere between these extremes. A natural next thought is that "A >> C" is true if and only if C is entailed not by A alone, but by A in conjunction with some other truths: these might include, for example, fundamental truths about the laws of nature. The problem with this suggestion is that no single set of fixed truths will do the job for all A and C. As Lewis (1973) notes, that would incorrectly include strengthening of the antecedent among the valid forms of inference for counterfactuals, yet counterfactuals are clearly *non-monotonic*. For example, from the fact that a particular match would light if it were struck, it does not follow that if the match were struck *and there were no oxygen present*, then it would light.

Counterfactuals are variably strict conditionals. In evaluating a counterfactual we see what is entailed by the antecedent along with some other truths, but the set of truths to be held fixed varies from case to case in a way that is determined by the antecedent A. In Goodman's phrase: we hold fixed everything that is *cotenable* with the truth of the antecedent. The threat of circularity should now be apparent, for what could it mean for a proposition to be "cotenable" with A, other than that the proposition is one that would still be true if A were true? One can hardly explain counterfactuals

in terms of the notion cotenability and then proceed to give a counterfactual analysis of "cotenable".

A way out of this circle was presented in the work of Stalnaker (1968) and Lewis (1973), inspired by the development of possible worlds semantics in modal logic. Central to the Stalnaker-Lewis approach is the assumption that possible worlds may be ordered with respect to their similarity to the actual world. Since the relation of comparative similarity to the actual world is assumed to be a weak order (i.e. a connected and transitive relation) we may usefully think of it as a relation of comparative "closeness". Call a world at which the proposition A is true an "A-world". Then the counterfactual "A >> C" is true if and only if C is true at the closest A-world to the actual world. If, following Lewis (1973), we want to allow that there may not be a single A-world, or even a set of A-worlds, closest to the actual world, the truth condition is more appropriately given like this:

> "A >> C" is true if and only if some (A&C)-world is more similar to the actual world than any (A&~C)-world is.

It is worth noting how much of the logic of counterfactuals follows simply from the assumption that the similarity relation is a weak order. We can see immediately, for example, the particular logical form of the fallacy of strengthening the antecedent. It is a fallacy analogous to the mistake made by someone who infers from "There is no hardware store in the closest town to here that has a bank" that "There is no hardware store in the closest town to here that has a bank and a restaurant". Further exercise of the analogy between similarity and closeness leads to the correct prediction that contraposition is also invalid, and that counterfactuals fail to be transitive. Appreciation of these points will be key to the discussion that follows.

While much of the logic of counterfactuals follows just from the assumption that any reasonable relation of comparative similarity will be a weak ordering, we will need to say something more substantial about the notion of similarity if our hope is for an account of counterfactuals on

which a theory of causation can be based. The problem is that similarity is multi-dimensional. If two worlds are each similar to the actual world in a different respect, which is most similar to the actual world all things considered?

We might in general be content to consign the weighing of respects of similarity and difference to the pragmatic rather than the semantic part of the theory. So we might say: in one context it is appropriate to assert that Caesar would have used the atom bomb had he been in command in Korea; in another that he would have used catapults (example from Quine REF). One of the contributors to this volume (Maslen, Chapter 14) is happy to allow that due to this feature of counterfactuals, the causal relation itself is context dependent; that whether or not it is correct to describe one event as a cause of another is relative to the contrast situation that one has in mind. But those who hope for a *context-independent* counterfactual account of causation must confront the issue of respects of similarity head on.

The issue arises most sharply in connection with temporal considerations. Causes precede their effects. A defender of a counterfactual theory of causation will likely think that this temporal asymmetry in causation reflects an asymmetry in counterfactual dependence: while the future depends counterfactually on the past and present, the past is not counterfactually dependent on the present and future.

Admittedly, we are sometimes willing to talk about how things would have been different in the past, had they been different now. ("If I had pulled the trigger just now, there would have to have been no-one in the line of fire. I'm not a homicidal maniac!") But as Lewis notes (1979a) such *backtracking conditionals* are typically "marked by a syntactic peculiarity"—note the "have to have been" construction in the above example. Backtracking conditionals will have to be set aside if we want to defend a counterfactual theory of causation. And they will have to be ruled out in some principled way. This provides a clue to the problem of weighing respects of similarity.

It is clear, as we noted above, that similarities with respect to the laws of nature (*nomological* similarities) should generally outweigh similarities with respect to accidental matters of fact in the all-things-considered evaluation of similarity. But apparently this cannot be an invariable rule.

That's one of the lessons of temporal asymmetry. For if the actual laws of nature are deterministic towards both past and future, then any world that shares these laws yet differs from the actual world in some particular present matter of fact will not only have a different future from the actual world, it will have a different past as well. We then have backtracking on a massive scale.

So we must allow that certain minor violations of the laws of nature ("small miracles") may be compensated for by agreement in past matters of particular fact. Yet it would hardly do to secure this agreement by stipulation. That would incorrectly (to our minds) deem backward causation an *a priori* impossibility. (Later in this section we will see that this issue is more complicated than the present remarks suggest, as accommodating backward causation turns out to be trickier than it might seem.) Nor, in the present context, is it an attractive option to appeal to *causal* notions—requiring, for example, that in a counterfactual situation in which some (actually occurring) event *c* does not occur, everything either causally antecedent to or causally independent of *c* be held fixed. Such a stipulation might guarantee the right truth values for the counterfactuals that feature in a counterfactual analysis of causation, but at too high a price in circularity.[2] The challenge for the Stalnaker-Lewis account of counterfactuals is to provide criteria for weighing respects of comparative similarity that yield temporal asymmetry without doing so by *fiat*.

The challenge has been pressed most vigorously in connection with what is known as the "future similarity objection" (see Fine 1975, and references in Lewis 1979a, p. 43 to another seven published versions of the objection). We may suppose the following counterfactual is true:

If Nixon had pressed the button, there would have been a nuclear holocaust.

But it seems as though according to the Stalnaker-Lewis theory it will come out false. For suppose a nuclear war never actually takes place; then any world in which no nuclear war takes place will be more similar to the actual world than any world in which a nuclear war does take place. A nuclear

war would, after all, make things very different from the way they actually are. Hence given any world in which the antecedent and consequent are both true, it is easier to imagine a closer world in which the antecedent is true and the consequent false. Just imagine some minor change that prevents the nuclear disaster from occurring.

In response Lewis says:

> The thing to do is not to start by deciding, once and for all, what we think about similarity of worlds, so that we can afterwards use these decisions to test [the analysis] ... Rather, we must use what we know about the truth and falsity of counterfactuals to see if we can find some sort of similarity relation—not necessarily the first one that springs to mind—that combines with [the analysis] to yield the proper truth conditions. (Lewis 1979a, p. 43)

By pursuing this strategy of "reverse engineering" the similarity ordering from our intuitions about particular conditionals, Lewis obtains a set of criteria (1979a, pp. 47-48) that, he argues, suffice to rule out backtracking in worlds like ours without making backward causation a conceptual impossibility. In brief, Lewis argues as follows: Take some event *c* that occurs at time t. Assuming determinism—and assuming that the world displays certain *de facto* temporal asymmetries that Lewis attempts to characterize (it is this assumption that is supposed to leave the door open for backward causation)—the only way to construct a counterfactual scenario in which (i) *c* does not occur at t, and (ii) there are no gratuitous departures from actuality, is to insert, shortly before t, a "divergence miracle": a localized violation of actual (though not counterfactual) law which throws history off course just enough to make it the case that *c* does not occur. Thus, we get a counterfactual situation in which history until shortly before t is exactly as it actually is, but whose history after t diverges substantially from actual history; moreover, only during a relatively short interval of time, and in a relatively localized region of space, do the events of this world fail to conform to the actual laws. Lewis argues that we cannot do things in reverse: there is no similar way to insert a single localized "reconvergence miracle" so as to arrive at a counterfactual scenario whose past before t differs from the actual past, but whose future from shortly after t onward is

---

[2] But see Woodward REF for a sophisticated development and defense of such an approach. Woodward grants the circularity, but argues persuasively that the resulting account of causation is nevertheless highly informative.

exactly like the actual future. It is in this way that his truth-conditions secure a non-backtracking reading of the conditional.

But worries remain on this score. For example, it is not clear whether future similarity should count for absolutely *nothing* in the ordering, or whether it should merely have relatively little weight. Here it helps to think about examples that are not deterministic, but which involve some genuine chance process. For simplicity, suppose that a certain coin toss is genuinely chancy. If I haven't tossed the coin yet, and the coin is fair, it seems wrong to say that "If I were to toss the coin, then it would land heads". It seems wrong to say this even if, as a matter of fact, the coin actually does land heads when I get around to tossing it. This suggests that worlds in which the coin will land tails when tossed are as close to the actual world as any world in which the coin will land heads. This means that there are non-actual worlds that are as similar to the actual world as the actual world is to itself. And of course that's just what one would expect to be the case if a certain respect of similarity or difference is to count for nothing.

But the interplay between counterfactuals and chance involves some subtlety. As Sid Morgenbesser reminds us, if I offer to bet you that the coin will land tails and you decline the wager, then I can say later, after the coin lands heads, "Look, you would have won!" How can the Stalnaker-Lewis account reconcile these conflicting intuitions?

A second and quite serious difficulty has recently been pressed by Adam Elga (2000).[3] Elga argues, very persuasively, that Lewis's attempt to provide truth-conditions for a non-backtracking reading of the counterfactual runs afoul of statistical mechanics. Recall the dialectic: For any event *c* occurring at time t in world w, there are worlds w' whose histories match that of w *exactly* until shortly before t, but which then exhibit a small "divergence miracle" that throws history off course just enough to guarantee *c*'s non-occurrence. But, Lewis claims (and *must* claim), provided w exhibits the appropriate *de facto* temporal asymmetries, there are *no* worlds w' in which (i) *c* does

---

[3] Albert 2000 also discusses the problem.

not occur; (ii) history from shortly *after* t on matches w's history exactly; (iii) this perfect match comes about as a result of a small "reconvergence miracle".

Elga shows that this last claim is false—at least, of many typical choices for *c*, in any world that displays the global entropic features of our own. Suppose, to borrow his example, that Suzy cracks an egg into a hot pan, and it cooks. Let *c* be her cracking of the egg. Pick a time t shortly after it has cooked. Now look at what happens from t *backward*: a cooked egg sits in a pan, gradually *un*cooking; it coalesces into a raw egg and leaps upward; a shell closes around it, forming a perfect seal. Surprisingly, the kind of time-reversibility that the laws of our world apparently exhibit allows that this entropy-diminishing process (flipped over in time, and run from past to future) is *possible* (see Albert 2000 for details); but they also guarantee that it is highly *unstable*, in the sense that a slight change in the initial conditions will upset the delicately timed processes that result in the uncooking and sealing of the egg (it is in establishing this result that statistical mechanics comes into play). Make such a slight change—say, by altering the positions of the molecules in one corner of the pan—and the egg will just sit there, slowly growing colder.

But, as Elga points out, such a slight change, inserted at t in the *actual* world, accomplishes exactly what is required of a "reconvergence miracle". That is, we construct a possible world by holding history beyond t fixed, making a slight, localized change in the state of the world at t, and running the laws backwards: in such a world, the egg is never cracked, and so *c* does not occur. The needed asymmetry between divergence and reconvergence miracles vanishes. There is real irony here: for Lewis closes his famous paper by observing, "I regret that I do not know how to connect the several asymmetries I have discussed and the famous asymmetry of entropy." (1979, p. 51)

Though serious, this problem should not be overstated. It is not just that some way might be found to amend Lewis's account in response to Elga's challenge (for example, one might construct standards of similarity that give significant weight to the global entropic profile of a world).[4] It is

---

[4] There is one other lesson from physics that is not so easily avoided, however. Here is an illustration: Suppose you have before you a glass of hot water. What would happen, were you to put an ice cube in it? We would like to answer: The ice cube would melt. But statistical mechanics, together with the time-reversibility of the fundamental

also that the philosopher who desires a non-backtracking reading of the counterfactual *solely* to serve the purposes of her counterfactual analysis of causation—and not also, as in Lewis (1979), to provide an account of the asymmetry between past and future—can fall back on a recipe for evaluating counterfactuals that guarantees, more or less by stipulation, that they will have the needed non-backtracking reading. As follows:

The first step is to require that causes must *precede* their effects. (More, in a moment, on whether this requirement carries too high a price.) Then the counterfactual analyst need only consider conditionals of the form "$\sim O(c) >> \sim O(e)$" where *c* and *e* both occur, and *c* before *e*. ("$O(c)$" is short for "*c* occurs", etc.) Suppose we are given such a conditional. To construct the relevant counterfactual situation, focus first on the time t at which *c* actually occurs. Consider the complete physical state of the world at that time. (Or: the complete state on some space-like hypersurface that includes the region of *c*'s occurrence.) Make changes to that state localized to the region in which *c* occurs, and just sufficient to make it the case that *c* does not occur. (These are the changes that Lewis's account would have brought about by a "miracle".) Evolve the new state forward in time in accordance with the actual laws, and check to see whether *e* occurs. If it does, the conditional is false; if it doesn't, true.

Matters might be more complicated: perhaps there are many equally minimal ways to change the given state, or an infinite succession of ever-more-minimal ways. If so, we employ the obvious sophistication: the conditional "$\sim O(c) >> \sim O(e)$" is true just in case *some* $\sim O(c)$-state that yields a $\sim O(e)$-future is "closer" to the actual state (i.e., involves a more minimal change from the actual state) than is any $\sim O(c)$-state that yields a $O(e)$-future.

---

dynamical laws that appear to govern our world, tells us that there would be some *extremely tiny* (but non-zero) probability that the ice cube would grow, while the water heated up. (Time-reversibility tells us that such an occurrence is nomologically possible, statistical mechanics that it is extremely unlikely.) If so, then the answer we would like to give is *false*—since it *might* be that the ice cube would *not* melt. As far as we know, there is no decent way to construct truth-conditions for the counterfactual that avoids this result. But the counterfactual analyst need not be greatly concerned: She need merely grant that her account of causation must be built upon conditionals of the form, "if c had not occurred, then the probability that e occurred would have been extremely close to zero." We will ignore this complication henceforth.

Notice that in following this recipe we have implicitly replaced a similarity metric on *worlds* with a similarity metric on *states*, or if you like, worlds-at-times. Not a problem—but it does make clear (what was perhaps obvious anyway) that the recipe is limited in scope, working best (perhaps: only) when the antecedent of the conditional is a proposition entirely about a particular time or narrow interval of time.[5] The counterfactual analyst should not worry: after all, she is doing the metaphysics of causation, and not the semantics of some fragment of English. And her counterfactual conditional evidently has enough in common with the ordinary language one still to deserve the name.

But two other problems plague the counterfactual analyst who takes our suggestion for responding to Elga's challenge. She has stipulated that causes must precede their effects. Has she not thereby ruled out—a priori—both simultaneous causation and backward causation? But both are surely possible: As to the first, we have, for example, the situation in which the particle's presence at a certain location in the force field at time t causes its acceleration with a certain magnitude at time t. As to the second, we have any of a number of seemingly perfectly consistent time-travel stories, e.g. the following: A billiard ball subject to no forces flies freely through space; call this ball "Bill". Off to the left is the Time Travel Exit Portal. Off to the right is the Time Travel Entrance Portal. At time 0, a billiard ball—call it "Sue"—emerges from the Exit Portal, flies towards Bill, and strikes it. Bill goes careening off toward the Entrance Portal, and crosses it at time 1.... And, of course, leaves the Exit Portal at time 0: Bill and Sue are one and the same.[6]

Examples such as these appear to depict genuine possibilities, but if causes must precede their effects then it seems that this appearance is illusory. In fact, however, we think matters are not nearly so simple. For reasons of space we will forego a full discussion; but we will try to say

---

[5] See Maudlin 2003 for discussion of how to extend this recipe to counterfactuals with more complicated antecedents.

[6] Observe that if backward causation is possible, and if causation is transitive, then simultaneous and indeed *self*-causation are possible: e.g., Bill/Sue's collision with itself causes that very same collision. Still, there is a distinct kind of simultaneous causation whose possibility cannot be secured in this way, where the causal relationship between *c* and its contemporaneous effect *e* is *direct*, involving the action of no causal intermediates. We intended the example of the particle in the force field to illustrate this sort of simultaneous causation.

enough to make it clear that these two problems should not prevent one from taking seriously the method we have suggested for guaranteeing the asymmetry of causation.

As to simultaneous causation, one promising strategy is to deny that the examples allegedly depicting it in fact succeed in doing so. Consider the example we gave, where the effect supposedly contemporaneous with its cause is the acceleration of a particle at time t. We suggest that the example gains its intuitive appeal only so long as we remain naive about the underlying physics. For what it *is* for a particle to be accelerating with a certain magnitude at time t is for its motion—*in an interval of time extending forward from t*—to display a certain character.[7] Put another way, it is *not* an intrinsic property of the state of the particle—or indeed of the world—as it is at time t that the particle has the given acceleration, then; rather, it is a *relational* property, where the relatum is the trajectory of the particle in the immediate future of t. We could properly insist, then, that when we loosely speak of the particle's position in the force field at time t causing its acceleration at that time, what we *mean* is that its position in the field causes a certain aspect of its *subsequent* motion. But then any conflict with the stipulation that causes must precede their effects disappears. Indeed, the example begins to look like the following example of "simultaneous causation": Say that a window is "doomed" at time t just in case it will shatter within ten seconds of t. Suzy throws a rock at a window at noon—instantaneously causing the event of its being doomed at noon. That is silly, the kind of example about which no competent philosopher would think twice. We think that careful reflection exposes the example of the particle in the force field as being equally silly. And, while we will not pursue the matter here, we suspect that other alleged cases of simultaneous causation can be given a similar exposé.

Backward causation requires, we think, a different treatment, for unlike some authors (e.g., Mellor REF), we do not wish to deny its possibility. Accordingly, we suggest a much more

---

[7] Here are the technical details: Let f(T) be a function that gives the particle's position for each time T (specified relative to some inertial frame of reference). We assume that the function is continuous, but not that it is everywhere differentiable. But suppose further that for some open interval extending forward from time T = t, the second derivative f''(T) is well-defined and continuous; then the particle's instantaneous acceleration at t, "taken from the

defensive maneuver, which is to point out that the need to accommodate backward causation provides, surprisingly, no reason whatsoever to return to Lewis's miracle-based account of the truth-conditions of the counterfactual.

To see why, return to our example of the collision of Bill/Sue with itself. Focus on a time moments before the ball crosses the Time Travel Entrance Portal. Let *c* be the event of its flying toward the Entrance, at that time; let *e* be the earlier event of its flying out of the Exit. Clearly, *c* causes *e*. But whether *e* counterfactually depends on *c* is not at all clear, and that is because it is not at all clear what is going on in a counterfactual situation in which *c* does not occur. Try first to construct this situation in a flat-footed way, by inserting a "divergence miracle" into the actual course of events just before *c* occurs, a miracle just sufficient to guarantee that *c* doesn't occur: you will quickly find that limiting yourself to just one such miracle cannot possibly do. For remember that the *rest* of what goes on—the totality of events that follow your miracle, and the totality of events that precede it—must *conform* to actual law. Suppose for illustration that your miracle makes the ball vanish, just before it reaches the Entrance. Then, by law, it cannot leave the Exit—not at time 0, nor at any other time. And there are no other balls in the picture. So *no* ball ever leaves the Exit. So no ball ever collides with Bill/Sue. So Bill/Sue continues on its original course. So, far from vanishing just before it reaches the Entrance at time 1, the ball is nowhere near it, then. Contradiction.

We could try to fix matters by inserting a second localized miracle: let the ball (miraculously) leave the Exit at time 0, even though it does not cross it at time 1. Then we have a consistent story, *most* of which is compatible with the actual laws: The ball is flying through space; a ball (miraculously) flies out of the Exit at time 0; the two collide, sending the original ball towards the Entrance; just before time 1, it (miraculously) vanishes. But *e* occurs, in this scenario. So if this is what would have happened, had *c* not occurred, then *e* does not depend on *c*. So the counterfactual analysis gets the causal structure of our case wrong.

---

future", can be defined as the limit as T approaches t "from above" of f''(T). (For a well-behaved particle, this will equal the limit "from below", and we can speak simply of the acceleration at t, simpliciter.)

We could also try to fix matters by insisting that the counterfactual situation in which *c* does not occur is simply one in which nothing ever leaves the Exit or crosses the Entrance: instead, the ball just flies through space, unmolested. So *e* does not occur; so—if this is what would have happened—*e* depends on *c*, after all. True, we do not reach this situation by inserting a miracle into the actual course of events. But Lewis's official criteria for similarity of worlds do not *require* miracles: one is supposed to introduce them only if doing so is necessary, in order to secure widespread and perfect match of particular fact.

The problem should now be obvious, for securing such perfect match is exactly what we accomplish by introducing a second miracle, whereby the ball leaves the Exit, even though it does not cross the Entrance. In fact, the first counterfactual situation we described matches the actual situation perfectly, *everywhere*, except for a tiny interval before time 1, in a small region of space near the Entrance. By Lewis's own criteria, it should therefore win out over the second situation. So the account of the truth conditions for counterfactuals in which these criteria feature does not, after all, seem so clearly preferable to the "asymmetry by fiat" account sketched above.

There is an important open question here, which we must simply note and pass over: How exactly can any sort of reductive account of causation—counterfactual or otherwise—accommodate backward causation? We strongly suspect that the best answers will end up treating backward causation as a kind of special case. If so, that may prove to be more good news for our suggested truth conditions for the counterfactual: for it may be to their credit that they allow for no straightforward treatment of backward causation, within the context of a counterfactual analysis. At any rate, those truth conditions evidently provide the ingredients necessary for pursuing a counterfactual account of less exotic varieties of causation. Let us now consider the prospects for such an account.

2. Reductionism about causation

It will be helpful to begin by situating counterfactual approaches to causation within the much larger constellation of philosophical accounts of causation quite generally. Probably the most useful

distinction to make at the outset is that between accounts that do, and accounts that do not, attempt to *reduce* causal facts to facts about what happens, together with facts about what the laws are that govern what happens. (We have a permissive sense of "what happens" in mind: it is to include facts about what objects exist where and when, and what categorical properties and relations they instantiate.) Immediately, two questions arise, which we will not pursue in any detail: First, can facts about the laws themselves be reduced to the totality of categorical facts? Some—notably, Lewis (e.g., 1983b)—will say "yes", others "no". We wish merely to note that with respect to the aims of the essays in this volume, almost nothing hangs on this dispute. (For example, Lewis's account of causation could be adopted wholesale by one who disagreed with him about whether the laws themselves reduce to categorical facts.) Second, is there in fact a legitimate distinction to be drawn between categorical and non-categorical facts about the world? Put another way, is it possible to specify, in *non-causal* terms, the facts about "what happens" that form part of our reduction base? The facts about the laws? We will simply proceed under the assumption of an affirmative answer; but see Shoemaker (1980) and Cartwright (1999) for influential statements of contrary views.

We will henceforth label "reductionist" any position according to which causal facts can be reduced to categorical plus nomological facts, and label "anti-reductionist" any position that denies this claim. Most of the essays in this volume are written from a reductionist perspective—or at any rate, from a perspective entirely congenial to reductionism. (Armstrong (chapter 19: "Going Through the Open Door Again") is a notable exception.) It is useful to add, however, that some anti-reductionists find strong motivation for their view in certain key thought experiments. Here, for example, is one provided by Michael Tooley (1990, italics in the original):

> Given [the assumption that there is nothing incoherent in the idea of an uncaused event], one can consider a world where objects sometimes acquire property Q without there being any cause of their doing so, and similarly for property R, and where, in addition, the following two statements are true:
>
> (1) It is a law that, for any object x, x's having property P for a temporal interval of length Δt either causes x to acquire property Q, or else causes x to acquire property R;
>
> (2) It can *never* be the case, for any object x, that x's having property P for a temporal interval of length Δt causes x to acquire both property Q *and* property R.

> Suppose, finally, that an object a in such a world, having had property P for the appropriate interval, acquires both Q and R. In view of the law described in statement (1), either the acquisition of Q was caused by the possession of P for the relevant interval, or else the acquisition of R was so caused. But, given statement (2), it cannot be the case that the possession of P for the relevant interval caused *both* the acquisition of Q *and* the acquisition of R. So once again, it must be the case that one of two causal states of affairs obtains, but the totality of facts concerning, first, the non-causal properties of, and relations between, events, secondly, what laws there are, and thirdly, the direction of causation in all potential causal processes, does not suffice to fix which causal state of affairs obtains.

In section 5 below, where we take up a number of tricky methodological issues, we will briefly touch upon the question of how much probative value thought experiments like this have (to anticipate: not much). Let us now proceed to draw further distinctions among reductionist approaches to causation.

Once again, a particular distinction stands out as especially helpful: it is that between what we will call "physical connection" accounts of causation and what we will call "nomological entailment" accounts. Examples will work better than definitions to give the idea.

The first kind of account, while reductionist, appeals to the fundamental laws only indirectly—typically, in the specification of some quantity whose "transfer" is thought to constitute the causal relation.[8] Thus, we might say, with Fair (1979), that causation consists in the transfer of *energy*. Or, with Ehring (1997), in the transfer of a *trope*. Or, with Dowe (1992) and Salmon (1994), in the transfer of some *conserved quantity*—where appeal is made to fundamental physics for an inventory of such quantities. This last example shows most clearly how the fundamental laws can have a place in such accounts.

We think such accounts suffer from a serious lack of ambition, twice over: First, however successful they are at limning the features of causation, as it relates events in the purely *microphysical* realm, there seems little hope that they can succeed in doing so in the messy macroscopic realm. Suzy kisses Billy, causing him to flush; are we to suppose that the causal relation between these two events is to be mapped out by looking at how energy, or some other

---

[8] Typically, but not always: Ducasse 1926, for example, takes causation to be essentially nothing more than spatiotemporal contiguity plus temporal priority. (That is, *c* causes *e* just in case *c* and *e* are spatiotemporally contiguous, and *c* precedes *e*.) We view this as a kind of limit case of physical connection accounts.

fundamental physical quantity, is transferred? Well, it might be romantic to say so. But not, we think, particularly enlightening.

Second, assuming as seems reasonable that it is a contingent matter what the fundamental laws are, physical connection accounts can not, because they are not designed to, tell us anything about causation as it might have been—in particular, as it is in worlds with laws very different from our own. That limitation seems not merely unfortunate but deeply misguided; for it seems clear that while the fundamental laws play an essential role in fixing the causal facts, they do not do so in so *specialized* a manner. Put another way, it seems both reasonable and worthwhile to try to specify the way in which the fundamental laws fix the causal facts in terms that *abstract away* from the gory details of those laws—thereby to produce an account that has a hope of proving not merely true, but necessarily so.

That is what nomological entailment accounts attempt to do. As a simple example—one that shows where the label "nomological entailment" comes from—consider the view that takes *c* to be a cause of *e* just in case it follows, from the proposition that *c* occurs, together with the proposition that encapsulates the fundamental laws, that *e* occurs (where the "following" could be spelled out in terms of deductive entailment, or in terms of some notion of metaphysical necessity: more on this in the next section). Or we might go the other way: *c* is a cause of *e* just in case it follows, from the proposition that *e* occurs, together with the proposition that encapsulates the fundamental laws, that *c* occurs. Or we might go both ways at once. (Lesson: do not take the "entailment" in question invariably to proceed from cause to effect.) While the defects in these accounts are perfectly obvious, it is *also* obvious that they do not suffer from the two kinds of constriction that made physical connection accounts seem so disappointing.

(The distinction between physical connection accounts and nomological entailment accounts does not exhaust the range of *reductionist* accounts of causation. Maudlin's "Causation, Counterfactuals and the Third Factor" (chapter 18) provides an excellent illustration of this point: He argues that our ability to apply any discriminating notion of cause is parasitic on our ability to analyze the *laws* governing the situation in question into, one the one hand, laws that specify how

things will behave if left undisturbed, and, on the other hand, laws that specify the exact consequences of particular disturbances.)

Of course, more sophisticated and correspondingly more interesting nomological entailment accounts are easy to find. For example, we might side with Mackie (1965): *c* is a cause of *e* iff, from the proposition that *c* occurs, together with the proposition that encapsulates the fundamental laws, together with some suitably chosen proposition about conditions that obtain at the time of *c*'s occurrence, it follows that *e* occurs, where the premise that *c* occurs is essential to the entailment. Or we might try building on the idea that *c* is a cause of *e* just in case the conditional probability that *e* occurs, given that *c* occurs, is greater than the unconditional probability that *e* occurs. (For a sophisticated version, see Eells, 1991.) Note that in calling this a "nomological entailment" account, we are, in the first place, assuming that the relevant probabilities are somehow determined by the laws (i.e., they are not subjective probabilities); and, in the second place, we are permissively counting probability-raising as a kind of "entailment". No matter: consider this a small sacrifice of accuracy for the sake of having a handy label.

More to the point, we might endorse the simplest counterfactual analysis: *c* is a cause of *e* iff, had *c* not occurred, *e* would not have occurred—where this entailment relation between the proposition that *c* does not occur and that *e* does not occur counts as "nomological" because of the central role played by the fundamental laws in fixing the truth-conditions of the counterfactual (cf. section 1). Or we might hold some more sophisticated counterfactual analysis that builds on the core idea contained in the simple version. Indeed, many philosophers hold out such firm hope for some sort of counterfactual analysis precisely because they are convinced, in the first place, that some sort of *nomological entailment* account must be correct; and, in the second place, that no rival to the counterfactual approach has a prayer. (See Lewis's "Causation as Influence", chapter 3 in this volume, for a particularly forceful defense of this motivation.) Be that as it may, in attempting such an analysis one should certainly come armed with a clear view of the major obstacles, a host of which have emerged over the course of the last 25 years. We will shortly survey the most important of them.

3. The original Lewis analysis

Let us first take stock of the principal *advantages* of the counterfactual approach. Focus on one of the simplest counterfactual analyses: Lewis's original (1973) analysis. He took causation to be the ancestral of counterfactual dependence: *c* causes *e* iff *c* and *e* both occur, *c* and *e* are distinct, and there is a (possibly empty) set of events $\{d_1, d_2,..., d_n\}$ such that if *c* had not occurred, d_1 would not have occurred; and if d_1 had not occurred, d_2 would not have occurred;... and if d_n had not occurred, *e* would not have occurred.

The analysis is attractively simple and extremely intuitive, given how obvious it seems that there must be *some* intimate connection between causation and counterfactual dependence. To be sure, the connection need not be explained by supposing that causation must be given a counterfactual analysis; see Maudlin's contribution (chapter 18) for a rival explanation. But the prima facie plausibility of a counterfactual analysis is sufficiently clear that no one can reasonably doubt that exploring its prospects is worthwhile.

Next, the account makes it exquisitely clear what the role is that laws of nature of play in fixing the causal facts—and moreover, makes it clear that this role does not require *too much* from the laws. Let us explain what we mean by this by contrasting the Lewis analysis with a Davidsonian account of causation. According to this latter kind of account, what distinguishes a true causal claim of the form "*c* is a cause of *e*" *as such* is that, when the names "*c*" and "*e*" are replaced by suitable uniquely identifying descriptions of the respective events, the resulting sentence can be seen to be a deductive consequence of some universal generalization or generalizations that hold with lawful necessity (together with additional premises specifying that the given events in fact occur). Morton White (1965, pp. 56-104) presents a clear statement of this sort of approach.[9]

---

[9] See also Davidson 1967. Note that Davidson does not treat the word "cause" in the way one might expect—namely, by providing it with an explicit definition that would allow for the deduction of the target sentence from sentences that did not themselves contain the word "cause". Rather, he includes the word "cause" in the statement of the laws themselves, a move that is both bizarre (given that the paradigm examples of laws provided by

This Davidsonian account faces a number of difficulties, but for our purposes the one to focus on is that it places severe constraints on an account of laws that can meet its needs. Are we to suppose, for example, that the lawful universal generalizations come directly from fundamental physics—so that the uniquely identifying descriptions have to be descriptions in the language of fundamental physics? Or is it rather that we can pick out higher-level laws tailored to any domain where we can find true causal claims? But then as good reductionists, we should want to know how these higher-level laws themselves reduce to facts about what happens, together with facts about the *fundamental* laws. Even if we are *not* good reductionists, we should want to know this—for unless we are told, we should view with serious skepticism the claim that there even *are* such suitable higher-level laws (enough of them, anyway).

Our simple counterfactual analysis elegantly avoids these issues. For as we in effect saw in section 1, all we need by way of an account of laws is an account of which worlds are *nomologically possible*; put another way, that is all that the laws need to contribute to the truth conditions for the relevant counterfactual conditionals. We do not need what the Davidsonian requires: for each causal relation, a law that will directly "cover" it, at least when the relata are described appropriately. (In hindsight, the crucial mistake the Davidsonian committed here was to require a deductive relation between *sentences* rather than a kind of metaphysical entailment relation between *propositions*.)

Next, if we remember that the semantics used for the kind of counterfactual conditional appropriate to counterfactual analyses will guarantee a *non-backtracking* reading of that conditional, then we can appreciate how the Lewis analysis neatly avoids three problems that were thorns in the side of earlier regularity analyses. First, in many cases where *c* is a cause of *e*, it will turn out that *e* is lawfully necessary and/or sufficient for *c* (at least, in the circumstances); on a number of regularity analyses, it will follow that *e* is a cause of *c*. More generally, those analyses tend to face a serious problem in explaining the asymmetry of the causal relation. One way to guarantee such

---

fundamental physics never make use of this notion) and that inexplicably ruins the chances for providing a reductive analysis of causation.

asymmetry is to insist that causes must precede their effects, thus piggybacking causal asymmetry on the asymmetry between past and future. But doing so seems rather hastily to rule out the possibility of either simultaneous or backward causation.

Still our discussion at the end of the last section may show that simultaneous causation need not be taken too seriously, and that backward causation is everyone's problem. So observe that this maneuver is of no help in solving the second problem, which arises from cases in which an event *c* causes two events *d* and *e* along different and independent causal routes. As an illustration, consider figure 1:

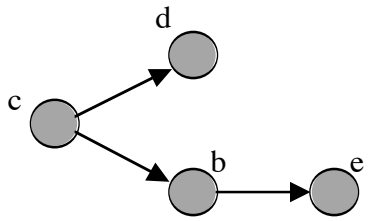

Figure 1

Here, **c** fires, sending a stimulatory signal to **d**, causing it to fire. (Circles represent neurons; arrows represent stimulatory connections; shading a circle represents the firing of the given neuron; the time order of events is left-to-right. In the text, **bold** letters name neurons, *italicized* letters the events of their firing.) In addition, **c**'s firing causes **e** to fire, by way of the intermediate neuron **b**. Now, given a suitable specification of the circumstances (e.g., which neurons are connected to which other neurons), it appears to follow from the fact that **d** fires, together with the laws, that **e** fires; and likewise vice versa. Moreover, the firing of **d** precedes the firing of **e**. But the one is not a cause of the other—trouble, once again, for any of a number of regularity analyses of causation. And stipulating that causes must precede their effects helps not at all.[10]

For the third problem, consider figure 2:

---

[10] But see Hall 2004 for an argument that a properly constructed regularity account can easily avoid this problem.

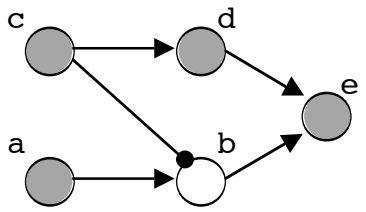

Figure 2

Here, the line with a blob at the end represents an inhibitory signal. Thus, **c** fires, stimulating **d** to fire, which in turn stimulates **e** to fire. Meanwhile, **a** fires, but the stimulatory signal it sends to **b** is blocked by an inhibitory signal from **c**, so **b** does not fire. Again, given a suitable specification of the circumstances, it follows from the fact that **a** fires, together with the laws, that **e** will fire. But to conclude from this fact, together, perhaps, with the fact that the firing of **a** precedes the firing of **e**, that **a** is a cause of **e** is simply to confuse *guaranteeing* an outcome with *causing* that outcome.

Thanks to the stipulation that the counterfactual conditional is to be given a non-backtracking reading, Lewis's analysis cleanly avoids each of these three problems (bracketing, for the moment, the reservations voiced in section 1 about some of the principal attempts to provide the needed truth conditions for non-backtracking counterfactuals). If the effect had not occurred, we are not entitled to conclude that it must have been the case that the cause did not occur: for to do so would be to impose the forbidden backtracking reading of the conditional. In figure 1, in evaluating what would have happened if **d** had not fired, we hold fixed the occurrence of other events contemporaneous with **d**—in particular, the firing of **b**—and hence conclude that it still would have been the case that **e** fired. Likewise in figure 2, if **d** had not fired, we do not conclude that it must (or even might) have been the case that **b** fired (on the basis that it must have been the case that **c** did not fire, and so must have been the case that no signal from **c** prevented **b** from firing), so that **e** would (or might) have fired all the same. Rather, we hold fixed the non-firing of **b**, and so endorse the conditional that if **d** had not fired, **e** would not have fired. (This last example, by the way, shows that the prohibition on backtracking is needed not merely to secure the sufficiency of the Lewis analysis, but also to secure its necessity: see Hall, "Two Concepts of Causation", chapter 9, pp. XXX-XXX.)

Thus, confronted with a suite of extraordinarily simple and straightforward examples that nevertheless proved stubbornly resistant to the regularity analyst's treatment, the counterfactual analysis appears to succeed quite effortlessly. What's more, the Lewis analysis appears to have the resources to handle quite naturally both causation by omission and causation under probabilistic laws. For the first, observe that we can take the failure of *c* to occur (or perhaps better: the failure of an event of type C to occur) to cause event *e* just in case, had *c* occurred (or: had an event of type C occurred), *e* would not have occurred. (Causation *of* omission can be given a similar treatment.) For the second, the counterfactual analyst can build on the idea that even under probabilistic laws, causes typically make a positive difference to the chances of their effects. That is, what distinguishes *c* as a cause of *e*, in the first instance, is that if *c* had not occurred, the chance of *e* would have been very much less than in fact it was (Lewis 1986b). Note that in the deterministic case, this condition reduces to the condition that had *c* not occurred, *e* would not have occurred (since the only chances available are 0 and 1). Many have therefore viewed its apparent ability to provide a uniform analysis of causation under both determinism and indeterminism to be a signal advantage of the counterfactual analysis.

The successes of the Lewis analysis are impressive; little wonder that Lewis's original paper had the influence it did. Still, further developments quickly showed the need to revise his proposal. We will focus now on the principal sources of pressure for revision, beginning with problems well-known for quite some time, and moving to problems that have only recently cropped up in the literature.

4. Problems, old and new

Some problems take the form of clean, simple counterexamples to the Lewis analysis. Others take the form of challenges to lay out its foundations more precisely. We will begin with three problems of the latter variety.

First, for reasons already addressed, one might doubt whether an adequate semantics for the counterfactual can be produced that will yield the needed non-backtracking reading of that

conditional. Observe that in so far as the Lewis analysis aims to be *reductive*, the project of giving such a semantics becomes especially difficult. We cannot say, for example, that the non-backtracking reading of "if *c* had not occurred, then *e* would not have occurred" is to be arrived at by considering a counterfactual situation in which *c* does not occur, but in which all of its *causes* do (and then checking to see whether, in this situation, *e* occurs).

Waiving this problem, one might worry, second, that we do not really have a handle on what sort of counterfactual situation we are supposed to be envisioning. That is, we are told that we must consider what would have happened if a particular event *c* had not occurred. To see how difficult it might be to do so, consider a simple example: Billy and Suzy converse, say, between noon and 12:30 one day. Let *c* be their conversation. What would have happened if *c*—that very conversation—had not occurred? Offhand, it is extremely difficult to say, simply because it is extremely difficult to discern where the boundary lies between those possible worlds in which that very conversation occurs, and those possible worlds in which it does not. After all, in order to map this boundary we should need answers to the following sorts of questions: Would *c* have occurred if Billy and Suzy had conversed, but about an entirely different topic? Would it have occurred if they had conversed about the same topic, but an hour later? A day later? A few yards away from where they in fact conversed? A hundred miles? Could *c* have involved entirely different people—or was it essential to it that it be a conversation between Billy and Suzy? Could their conversation *c* have been, in fact, a race they were running together? Presumably not—but *why* not? To answer this and all the other questions, we should need an account of which features of it are essential to their conversation, and which accidental. Without such an account, it seems that we are at sea with respect to evaluating the very conditionals that form the heart of the Lewis analysis.

Unfortunately, this problem has tended to get overlooked in the literature. (See Bennett 1988 for a notable and welcome exception.) Part of the reason, no doubt, is that in the sorts of simple "neuron world" examples that often drive discussion in the literature, it is clear enough what counterfactual situation the author has in mind: letting *c* be the firing of some particular neuron, the counterfactual situation in which *c* does not occur is a situation in which that neuron doesn't fire,

but rather remains idle. More generally, if event *c* consists in the F-ing of some particular x at some time t, then we might naturally construe the counterfactual situation in which *c* does not occur to be, simply, a situation in which x does not F at time t. That does not yet solve the problem, for it is not necessarily clear how we decide what x would be doing *instead*. But here too, there might be a natural choice. Thus, if a neuron were not firing at a certain time, it would be idle instead. (It would not be glowing pink, or have turned into a dove, or have disappeared, etc.) Similarly, if Billy and Suzy had not conversed between noon and 12:30, they would have done some other activity normal for them instead (played, read books, sat quietly, etc.).

But to say that some such rough and ready rules guide our understanding of the counterfactuals that appear in the Lewis analysis does not *help* that analysis, so much as make more vivid one crucial shortcoming in its foundations (namely, that no clear and precise rules have been laid down for evaluating the counterfactuals). Unless, of course, the analysis is amended so as to explicitly incorporate these rules (in, one would hope, a much less rough and ready form). Note, finally, that while Lewis is aware of this problem, his own remarks are not particularly helpful. Here is what he says, in "Causation as Influence", chapter 3 (p. XXX, italics added):

> What is the closest way to actuality for C not to occur? —It is for C to be replaced by a very similar event, one which is almost but not quite C, one that is just barely over the border that divides versions of C itself from its nearest alternatives. But if C is taken to be fairly fragile, then if C had not occurred and almost-C had occurred instead, very likely the effects of almost-C would have been much the same as the actual effects of C. So our causal counterfactual will not mean what we thought it meant, and it may well not have the truth value we thought it had. When asked to suppose counterfactually that C does not occur, we don't really look for the very closest possible world where C's conditions of occurrence are not quite satisfied. Rather, we imagine that *C is completely and cleanly excised from history*, leaving behind no fragment or approximation of itself.

We leave it to the reader to try to figure out what it means to "cleanly excise" an event. (Fill the spacetime region in which it occurs with vacuum?) That an answer will likely prove elusive reinforces the suspicion that this issue needs to be taken more seriously by advocates of counterfactual analyses.

A third and related challenge for the counterfactual analyst—and, to be fair, for almost anyone who wishes to come up with a philosophical account of causation that treats events as the primary

causal relata—is to provide some sort of account of what events *are*. Without such an account it may be difficult to fend off various annoying counterexamples. For instance:

Suppose that Suzy shuts the door, and in fact slams 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, Suzy had shut the door softly). But if she hadn't shut the door, she 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 Suzy sends an email message exactly two days before t. Let *c* be such an instance, and let *e* be the reply (from Billy, of course), occurring at, say, a time one day after Suzy 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 backward 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—and so are not eligible to stand in a causal relationship—because of their logical relationship (some would prefer: because there is only *one* event, differently described). 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 genuine events

cannot have such extrinsic "essences". 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.

Still, it is hardly a peculiarity of the counterfactual analysis that it needs this sort of supplementing. Suppose, following 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 we 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*. (For relevant discussion, see Kim 1971.) Contrast the earlier problem about evaluating the relevant counterfactuals: that problem *does*, off-hand, seem idiosyncratic to counterfactual analyses.

Let us turn now to the simple counterexamples, both old and new, that have forced those pursuing a counterfactual analysis to add successive refinements to Lewis's original version. It will in fact be helpful to rewind slightly, and consider a version even simpler than Lewis's: namely, the analysis that says that *c* is a cause of *e* iff *c* and *e* both occur, but had *c* not occurred, *e* would not have occurred. Figure 2 already showed why this analysis won't work; for if **c** had not fired, **e** would have fired all the same, as a result of the "backup" process initiated by the firing of **a**. Hence Lewis's decision to take causation to be the *ancestral* of counterfactual dependence: for while the firing of **e** does not depend on the firing of **c**, we can trace a chain of step-wise dependence via the firing of **d**.

But there are two problems with the simple fix. The first is that it automatically commits the analysis to the view that causation is transitive—a view which, however intuitive, has recently come under suspicion, thanks to a range of purported counterexamples (see section 6 below). These counterexamples can, perhaps, be resisted. But a second and more serious problem is that there are

simple ways to tweak the example depicted in figure 2 that render Lewis's fix inapplicable. For instance, consider figure 3:

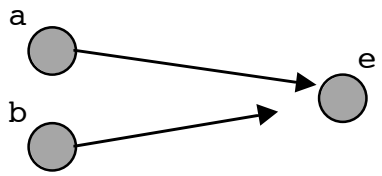

Figure 3

Here, **a** and **b** fire at the same time, both sending stimulatory signals to **e**. However, the signal from **b** is a bit slower (a fact represented by drawing its arrow so that it does not quite connect with **e**). Hence the signal from **a**, arriving first, is what causes **e** to fire. Countless real-world examples duplicate this structure, e.g. the following: Billy and Suzy both throw rocks at a window; each throw would be enough, by itself, to break the window; but Suzy's rock gets there first, hence it is her throw, and not Billy's, that is a cause of the subsequent shattering.

Pick the firing of **a**—or indeed, any event in the sequence constituting the passage of the stimulatory signal from **a** to **e**—and consider what would have happened if it had not occurred. Answer: **e** would have fired all the same, as a result of the signal from **b**. There is no hope, then, of tracing a chain of step-wise dependence from the firing of **a** to the firing of **e**.

The literature has dubbed this the problem of "late pre-emption" (as contrasted with the problem of "early pre-emption" exhibited, for example, by figure 2). A natural enough response to it is to focus attention on the *timing* of **e**'s firing, pointing out that if **a** had not fired, **e** would have fired moments later, whereas if **b** had not fired, **e** would have fired at exactly the same time. The literature has seen a number of attempts to leverage this feature of the case into a solution to the problem of late pre-emption. We won't try to review these here (but see the extended version of Lewis's "Causation as Influence", chapter 3, for a valuable and comprehensive discussion); we do, however, want to suggest that the pursuit of this kind of solution may be misguided. Suppose, for example, that the signal from **a** exerts a retarding force on the signal from **b**, slowing it down slightly. Suppose that the exact strength of this force is such that, had **a** not fired, the signal from **b** would have arrived at **e** at exactly the same time that the signal from **a** in fact arrives. Then not only

does the firing of **e** not depend on the firing of **a**, but nothing *about* the firing of **e**—not its timing, nor any other feature of its manner of occurrence—depends on the firing of **a**. (We can further suppose that the retarding force is of just the right strength that, for exactly the same reason, *e* exhibits a total lack of dependence on each of the events that consist in the passage of the stimulatory signal from **a** to **e**.)

At any rate, it is safe to say that the problem of late pre-emption has proved quite stubborn, and has provoked vigorous discussion in the literature. Several of the contributions in this volume give up-to-the-minute treatments of it.

There is another way to tweak the example of figure 2 so as to foil Lewis's original fix. It involves the controversial assumption that there can be action at a temporal distance. Suppose, as in figure 4, that one effect of **c**'s firing is to cut off the stimulatory signal from **a**, as before; but suppose further that when **c**'s firing causes **e** to fire, it does so not by way of any intermediate events, but rather by acting *directly* on **e** (albeit at a spatial and temporal distance; we represent this direct action by a thick, shaded arrow):

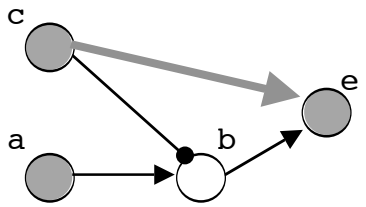

Figure 4

Once again, the situation provides no materials for tracing a chain of dependence from *c* to *e*. We leave it to the reader to consider whether this sort of example is more stubbornly resistant to treatment than the kind of late pre-emption depicted in figure 3.

We noted above that one of the features of the Lewis analysis that made it appear particularly promising was that it could give a clean and uniform account not only of ordinary causation under determinism, but of causation by omission, and of causation under indeterminism. But serious complaints have been raised with respect to these latter two points. About causation by

omission—and, more generally, other kinds of causal fact that seem to involve omissions (such as prevention)—two problems call for particular attention.

First, the typical counterfactual condition deemed necessary and sufficient for causation of event *e* by the omission of an event of type C—namely, that no event of type C occurs, that *e* occurs, and that had an event of type C occurred, *e* would not have occurred—has struck many as wildly permissive. Suzy and Billy plan to meet for lunch, but Billy doesn't show up; consequently, Suzy becomes sad. So far, no problem: Billy does not show up (i.e., no event of the Billy-showing-up type occurs); Suzy becomes sad; and, if Billy had shown up (i.e., if an event of the omitted type had occurred), Suzy would not have become sad. But something *else* that doesn't happen is Bill Gates's showing up to present Suzy with a $10 million check. And, if he had, she would certainly not have become sad (she doesn't like Billy *that* much). Are we to conclude that his failure to show up with the check likewise caused her to become sad? Once you see the trick, the embarrassing examples quickly proliferate. For in general, for any event *e* there will typically be a vast multitude of ways that it could, in principle (and compatibly with the laws), have been prevented. Are we to say that for each such would-be preventer, its failure to occur was among *e*'s causes? The counterfactual analyst seems committed to a "yes" answer here; but again, to many, that answer seems absurd.

So posed, the challenge invites the response that it is really everyone's problem—that is, that given modest assumptions, it follows that there is either a whole lot of causation by omission (much more than we would ordinarily feel comfortable admitting), or that there is none at all. For example, what principled basis is there for distinguishing Billy's failure to show up for lunch from Bill's failure to show up with a check? That Billy promised to, whereas Bill did not? But how could that sort of normative consideration have any place in a proper account of *causation*? Beebee's contribution (chapter 11: "Causing and Nothingness") presses this argument, and both she and Lewis (chapter 3: "Causation as Influence", and especially chapter 10: "Void and Object") endorse this disjunctive "either an awful lot or none at all" conclusion, albeit for very different purposes. (See also McGrath REF for a sophisticated treatment of this issue.)

For the second problem, consider how a counterfactual analyst might analyze the notion of *prevention*. A natural approach is to say that event *c* prevents event *e* from occurring (or perhaps better: prevents an event of type E from occurring) just in case *c* occurs, *e* does not (or: no event of type E does), and if *c* had not occurred, *e* (or: an event of type E) would have occurred. Thus, Billy throws a baseball at the window, and Suzy prevents the window from breaking by catching it: that is, her catch occurs, the window does not break (no event of the window-breaking type occurs), but if she hadn't caught the baseball, the window would have broken.

So far, so good. But now tweak the example so that it becomes an instance of what is commonly called "preemptive prevention": Suzy's friend Sally stands behind her, ready and able to catch the ball if Suzy does not. In fact, because Suzy catches the ball, Sally doesn't have to do anything. With that change, we have lost the dependence of the window's failure to break on Suzy's catch; but for many, the intuitive verdict that Suzy's catch prevents the window from breaking still holds. After all, Sally just stood there, doing nothing—and since *something* obviously prevented the window from breaking, it had to be Suzy (or rather, her catch). On the other hand, many find that Sally's presence undermines the original verdict—or if Sally is not enough, then substitute in her place the presence of a high, thick, sturdy brick wall. After all, how can Suzy get credit for preventing the window from breaking, when it was never in any danger from Billy's throw? The issues here are subtle, for it is a difficult matter to tease out exactly what is guiding our intuitions in such cases. (For discussion, see the contributions by Maudlin (chapter 18: "Causation, Counterfactuals, and the Third Factor") and Collins (chapter 4: "Preemptive Prevention"); we will briefly return to this kind of case in section 5, below.)

Still, if, with Sally present, the intuitive verdict that Suzy's catch prevents the window's breaking *stands*, then that is trouble for the counterfactual analysis. For it is not at all clear how any of the standard tools for handling *ordinary* cases of preemption can be applied to get Suzy's catch—and *only* Suzy's catch (not also Sally's presence, or the presence of the wall, etc.)—to connect up in the right way with the window's failure to break.

Next, a serious problem has also emerged for the standard counterfactual treatment of causation under probabilistic laws. Recall the key idea: *c* is a cause of *e* if *c* and *e* occur, but if *c* had not occurred, then the chance of *e*'s occurring would have been very much lower than in fact it was. (We then add various refinements to extend that sufficient condition for causation into a full-blown analysis of causation.) As an example, consider figure 5:

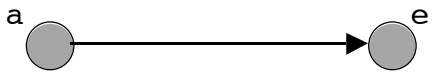

Figure 5

Here, **a** fires, sending a stimulatory signal to **e**, which fires. Let the laws state that **e** is certain to fire if it receives such a signal, but has a tiny chance—say, 0.01—of firing spontaneously, even if it receives no signal. So far, so good: our sufficient condition correctly classifies **a**'s firing as a cause of **e**'s firing, in this case.

(Some would object, wanting to insist that it remains up in the air whether **e**'s firing was caused by the signal from **a**, or rather was one of those rare instances of a spontaneous firing. But no reductionist can endorse this objection. For what difference—either in categorical facts, or in the laws—could distinguish the two situations? Note, in this regard, the close parallel between the example depicted in figure 5 and Tooley's anti-reductionist thought experiment discussed above.)

But now consider a variant: We allow that there is some small chance—again, say, 0.01—that the signal from **a** will die out shortly before it reaches **e**. Suppose, as in figure 6, that it does so:

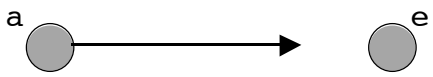

Figure 6

But **e** fires all the same. Then the trouble is that even though it is clear that **a**'s firing is not a cause of **e**'s firing, it still seems that the sufficient condition is met: **a** fires, **e** fires, and if **a** had not fired, then the chance of **e**'s firing would have been very much less (0.01 instead of slightly greater than 0.99).

A possible fix is to be more careful about the time at which we evaluate the chance of **e**'s firing. For in the depicted scenario, it is true that immediately after **a** fires, the chance that **e** will fire when it does is close to 1. But, letting time t be a time immediately before **e** fires, but *after* the signal from **a** has fizzled out, it is also true that the chance, at t, of **e**'s firing is a mere 0.01—exactly what it would have been, if **a** had not fired. So the t-chance of **e**'s firing does not counterfactually depend on **a**'s firing. It remains to be seen whether this fix could be extended into a proper counterfactual treatment of causation under probabilistic laws; in particular, it is not clear how the original sufficient condition should be modified, in order to be able to correctly classify some cases of causation as such, and not merely avoid misclassifying some cases of non-causation. See the contributions by Kvart (chapter 15: "Causation: Probabilistic and Counterfactual Analyses"), Ramachandran (chapter 16: "A Counterfactual Analysis of Indeterministic Causation"), and Hitchcock (chapter 17: "Do All and Only Causes Raise the Probabilities of Effects?") for highly sophisticated treatments of causation under indeterminism.

Turn now to more novel counterexamples. The first of these has come to be called "trumping" pre-emption. As an illustration, consider figure 7:

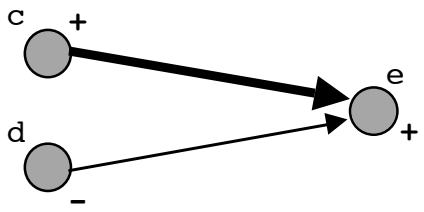

Figure 7

Let us suppose, here, that neurons can fire in different intensities, and with different polarities ("positive" and "negative"). Suppose further that a neuron that receives a single stimulatory signal fires in the polarity of that signal (although with an intensity independent of the intensity of the stimulating signal). Finally, when, as here, a neuron receives two stimulatory signals simultaneously, it fires in the polarity of the more intense signal: in this case, the signal from **c** (its greater intensity being represented by drawing a thicker arrow). No problem yet—or at least, no very serious problem: for as long as we are willing to assert that the positive firing of **e** that in fact

occurs would not have occurred if **e** had fired negatively, then we can safely conclude that the firing of **e** depends on the firing of **c**.

But now tweak the example in the obvious way, so that **c** and **d** fire with the same (positive) polarity; but as before, let **c** fire more intensely:

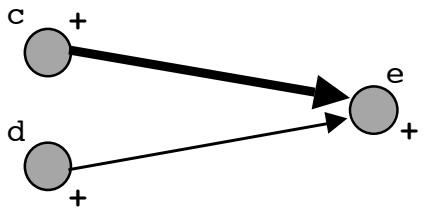

Figure 8

Many consider it just as intuitive that in figure 8 it is *c* alone—and not *d*—that is a cause of *e*. But if **c** had not fired, **e** would have fired in exactly the same manner; and the same is true if we focus not upon **c**, but rather on any event in the passage of the stimulatory signal from **c** to **e**. (For detailed discussion—which shows how wide is the range of counterfactual analyses that this example threatens—see Schaffer, "Trumping Pre-emption", chapter 2 in this volume.)

For the next novel counterexample, consider "double prevention". We will borrow Hall's example:

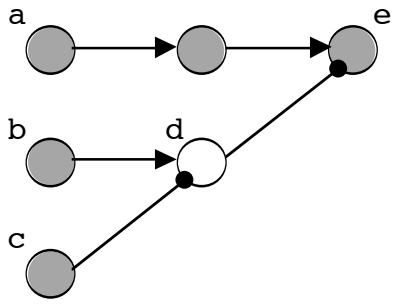

Figure 9

In figure 9, neurons **a**, **b**, and **c** all fire simultaneously; **c**'s firing prevents **d** from firing. Figure 10 depicts what would have happened, had **c** not fired:

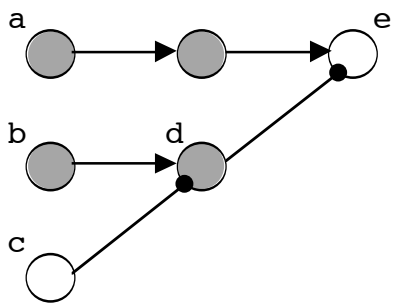

Figure 10

Evidently, **e**'s firing counterfactually depends on **c**'s firing; for that reason, almost every counterfactual analysis will count *c* a cause of *e*.

Some would consider that result a counterexample as it stands, since it strikes them as intuitively wrong that *c* is a cause of *e*, here—the idea being that *c* is not connected up to *e* in the right sort of way. Hall (in "Two Concepts of Causation", chapter 9 in this volume) argues in favor of this intuition by drawing what he takes to be three damaging conclusions from the claim that *c* is a cause of *e*: first, that the counterfactual analysis will lose the ability to distinguish genuine from ersatz action at a distance (for *c* and *e* are not connected by a spatiotemporally continuous causal chain); second, that it will be forced to give up the idea that, roughly, the causal structure of a process is intrinsic to that process (for the fact that *e* depends on *c* can easily be made to be extrinsic to the processes depicted); and third, that it will be forced to deny that causation is transitive (for reasons we will take up in section 6, below). Whether this last consequence is really so devastating is controversial, for the recent literature has also seen a spate of counterexamples to transitivity itself—and not merely to the conjunction of transitivity with one or another counterfactual analysis.

Let us now turn our attention away from the prospects and problems for a successful counterfactual analysis of causation, and consider a number of broader issues in the philosophy of causation. As will be apparent, tight connections remain: for almost every topic we will discuss in the sections ahead has been illuminated in myriad and interesting ways by work in the counterfactual tradition. We begin with a brief look at

5. Methodology

Work on philosophy of causation is, not surprisingly, heavily driven by intuitions about cases. Standard procedure often seems to be the following: A philosopher proposes a new analysis of causation, showing how it delivers the intuitively correct results about a wide range of cases. But then novel cases are proposed, and intuitions about them exhibited which run counter to the given theory—at which point, either refinements are added to accommodate the recalcitrant "data", or it's back to the drawing board.

It's worth taking time out to consider some methodological questions concerning this procedure. That is what we will do in this section.

One obvious and extremely difficult methodological question is this: What sort of project are philosophers of causation engaged in, that consulting intuitions about cases would be an appropriate way to pursue it? We won't have much to say about that question, not merely because it is so difficult, but also because it does not have much to do with work on causation *per se*. The methodology in question is so widespread in philosophy that the question should really be asked of the field as a whole. Still, it will be useful to distinguish two quite different aims that a philosopher working on causation might have. On the one hand, she might, like Lewis, take herself to be providing a good old-fashioned conceptual analysis of causation—a detailed explanation, that is, of how our ordinary concept works (see especially section 1.1 of Lewis's "Causation as Influence"). (Note that in doing so, she need not take on the extra commitment to a *reductive* account—reductionism is an optional, if natural and attractive, extra.) On the other hand, she might view her account as at least partially stipulative—that is, as providing a cleaned up, sanitized version of some causal concept that, while it may not track our ordinary notion of causation very precisely, nevertheless can plausibly be argued to serve some useful theoretical purpose. (Or one might fall somewhere in between: certainly, the distinction admits of plenty of gradations.)

Obviously, someone who pursues this latter aim ought to say at some point what such purposes might be. But we think that she is no under no obligation to make this clear at the outset. On the contrary, it strikes us as a perfectly appropriate strategy for a philosopher working on causation to

try to come up with a clean, elegant, theoretically attractive account of causation (or of some causal concept), in the reasonable expectation that such an account will serve some, possibly as-yet undisclosed, philosophical or perhaps even scientific purpose. And that expectation can be reasonable even if it is clear—because of recalcitrant intuitions about certain cases—that the account does not earn its keep by providing an explanation of how our *ordinary* concept works. This tolerant methodological perspective gains further support from the observation that causal concepts are used all over the place, both in philosophy and in the sciences; indeed, just sticking to philosophy, it often seems that for any philosophically interesting X, there is at least one "causal theory of X" on the market. It would be hasty to assume that the causal concept or concepts at work in any such theory is just our plain old ordinary one. In short, then, there is good reason to think that there is plenty of work available for those philosophers of causation who take themselves to be in the business of, as it were, "conceptual synthesis", rather than old-fashioned conceptual analysis.

Now, it is clear enough—at least, for present purposes—why someone interested in providing a conceptual analysis of our ordinary notion of causation should attend carefully to intuitions about cases. What we wish to emphasize is that even someone interested in "synthesizing" a new and potentially useful causal concept needs to heed these intuitions, else she risks cutting her project free of any firm mooring. More specifically, a reasonable and cautious approach for her to take is to treat intuitions about cases as providing a guide to where interesting causal concepts might be found. Thus, while her account can selectively diverge from these intuitions, provided there are principled reasons for doing so, it should not diverge from them wholesale.

It is particularly clear that one should pay attention to intuitions about cases if the stipulative element in one's analysis only consists in resolving various ambiguities or indeterminacies in our ordinary notion of "cause", or in some other way refining or precisifying that notion. For example, one might think that our ordinary notion is context-sensitive in its application, in various ways; and one might be interested in looking for a closely related notion from which such context-sensitivity has been expunged. Or one might suspect that our ordinary notion of causation involves an

unsupportable element of anti-reductionism—as witness the intuitive pull (such as it is) of thought-experiments like Tooley's; one might therefore try to produce a sanitized, reductionist-friendly surrogate for our ordinary notion. More provocatively, one might hold with Hall ("Two Concepts of Causation") that there is no way to provide a univocal analysis of our ordinary notion of causation, and that therefore the best thing for a philosopher to do is to break it up into two or more distinct concepts—distinct, at least, in the sense that they deserve radically different analyses. The list goes on: indeed, even a cursory survey of work in philosophy (as well as in the conceptual foundations of other disciplines) will reveal numerous precedents for the view that there can be legitimate philosophical studies that do not count as conceptual analysis of some ordinary concept, but that still require close attention to intuitions involving the application of such a concept.

It is not always clear in the literature what a given author's aim is: conceptual analysis, conceptual synthesis, or perhaps something else entirely. But it is important to be aware of the options, for otherwise it can be quite difficult to assess the cogency of certain typical responses philosophers make when confronted with recalcitrant intuitions. Above, we observed that *one* typical response—the most obvious one—is to go back and revise the theory. But that is hardly the only response. In addition, philosophers will often try to argue that the recalcitrant intuitions need not be respected. There are a number of strategies for doing so, and philosophers are not always careful to make explicit which strategy or combination of strategies they are pursuing. So we think it useful to list, and briefly comment on, some of the most prominent ones. As we will see, in several cases the effectiveness of the strategy depends crucially on whether or not the theory being defended is offered as a conceptual analysis of our ordinary concept of causation.

Suppose, then, that we have some philosophical theory of causation T that issues some verdict about some case, and that ordinary intuition about the case runs counter to the verdict. Then there are at least eight distinct ways that we might try to fend off the recalcitrant intuition.

First way: The intuitions in question are really not so firm—so the case is an example of "spoils to the victor".

Lewis has provided a particularly nice statement of the idea behind this strategy:

> When common sense delivers a firm and uncontroversial answer about a not-too-far-fetched case, theory had better agree. If an analysis of causation does not deliver the common-sense answer, that is bad trouble. But when common sense falls into indecision or controversy, or when it is reasonable to suspect that far-fetched cases are being judged by false analogy to commonplace ones, then theory may safely say what it likes. Such cases can be left as spoils to the victor, in D. M. Armstrong's phrase. We can reasonably accept as true whatever answer comes from the analysis that does best on the clearer cases. (Lewis 1986b, p. 194)

As an example, consider a case of perfectly symmetrical overdetermination: Suzy and Billy both throw rocks at a window; the rocks strike at the same time, with exactly the same force; the window shatters. Furthermore, each rock strikes with sufficient force to shatter the window all by itself. There is some intuition here that both Suzy's and Billy's throws are causes of the shattering. But this intuition is far from firm. Hence, when a counterfactual analysis says—as it very likely will—that neither Suzy's nor Billy's throw was a cause of the shattering (because it depended on neither, etc.), then that verdict can perhaps be considered one of the "spoils" to which the analysis is entitled (provided, of course, that it emerges victorious in its competition with its rivals).

But Lewis's claim makes little sense if one is pursuing a genuine *conceptual analysis* of our ordinary notion of "cause". For a successful such analysis will show how our concept works; so if our concept goes indeterminate in a particular case, then the analysis had better show that and why it does so. (Obviously, such an expectation is *not* appropriate if one's aim is to refine, precisify, or in some other way modify our ordinary conception of causation.) For comparison, suppose that we offered an analysis of "bald" according to which anyone with less than 500 hairs on his or her head counted as bald, and anyone with 500 or more hairs counted as not bald. Never mind the many other reasons why the analysis is inadequate (it matters how the hair is distributed, etc.): the one to focus on is that leaves it completely opaque why there should ever be a borderline case of "bald". That would be reason enough to count it a failure, *if* we had offered it as an account of our *ordinary notion* of "bald".[11]

---

[11] Given that Lewis quite explicitly takes himself to be providing a conceptual analysis of causation, it is therefore mysterious to us why he continues thus: "It would be still better, however, if theory itself went indecisive about hard cases. If an analysis says that the answer for some hard case depends on underdescribed details, or on the resolution of some sort of vagueness, that would explain nicely why common sense comes out indecisive." Lewis

At any rate, the example of symmetrical overdetermination also illustrates a very important if unrelated point. For while we might happily reject whatever intuition there is to the effect that each of Billy's and Suzy's throws is a cause of the shattering, we should *not* be happy with the conclusion that the shattering lacks causes altogether—and on the face of it, counterfactual analyses threaten to yield this conclusion. (Some, e.g. Lewis 1986b, try to defend the view that the *mereological sum* of the two throws causes the shattering.) Our point is not to press this problem on those analyses, but simply to observe that, in order to handle a difficult case, it may not always be enough simply to argue away the recalcitrant intuitions about it.

Second way: The case is misleading, because it is easily confused with other, unproblematic cases.

As an example of a case where this strategy might apply, consider the instance of action at a temporal distance depicted in figure 4. One might try to argue that in forming intuitive judgments about this case—in particular, the intuitive judgment that *c* is a cause of *e*—it is extremely hard not to let one's intuitions be guided by the thought that there is some faint, indetectable signal transmitted from **c** to **e** that is responsible for establishing the causal connection between the firings of the two neurons.

Or again, consider the example of trumping depicted in figure 8. One might argue that, in judging any such case, it is very hard not to presuppose that the trumping pattern is present in virtue of some unspecified and hidden physical mechanisms. (And when that is the case, examples of trumping become much less problematic; for further discussion see Schaffer, "Trumping Pre-emption", chapter 2.)

Third way: The case is misleading, because misleadingly presented.

As an example, let us return to Tooley's anti-reductionist argument. Observe that in setting up that argument, he invites us to consider a possible world in which it is a law that when an object has property P, its possession of this property causes it to acquire either property Q or property R—but

---

nowhere explains why he thinks it would merely be "better", as opposed to crucial, to explain "why common sense comes out indecisive" when it does.

never causes it to acquire both. Note that the word "causes" is crucial to the specification of the content of the laws: for we are *also* invited to believe that in this possible world an object can, consistently with the laws, acquire both Q and R after coming to possess P. Thus we cannot consistently replace "causes" by "is followed by". But to describe the relevant laws in this way prejudices matters enormously. After all, a good reductionist about causation will hold that the content of the fundamental laws can, in principle, be specified in *non-causal* terms—indeed, can perhaps be identified simply with a set of nomologically possible worlds. (And it is well to remember that in so far as we wish to look to the physics of our day for paradigm examples of fundamental laws, this reductionist position is by far the more sensible: Schrödinger's equation, for example, makes no mention of *causal* relations between physical states.) Tooley's thought experiment presupposes a very different conception of laws; it seems to us that the reductionist is well within her rights to reject this conception. She should insist that Tooley has either mischaracterized the content of the laws, or has not succeeded in describing a genuine possible world.

Fourth way: The example is misleading, because it naturally draws our attention to other, related concepts; we therefore mistake the conditions appropriate for applying these other concepts with the conditions appropriate for applying the concept of causation.

Beebee's contribution ("Causing and Nothingness", chapter 11) provides an excellent case study of this strategy at work. She argues that there is no such thing as causation by omission, but that we can perfectly well *explain* the occurrence of some ordinary event by citing the failure of some other kind of event to occur, where an occurrence of that kind of event would have prevented the explanandum event. Beebee goes on to argue that the reason that we often mistakenly think that there is causation by omission is that it is very easy for us to confuse causation with explanation.

As another example—one where this strategy should have been applied, but wasn't—consider the following passage from 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.

The case is a bad one, because it is very easy to confuse the question of whether the doctor's action is a *cause* of the patient's death with the question of whether, by so acting, the doctor *killed* the patient. The answer to the latter question is quite clearly "no"; confuse it with the former, and you will think that the answer to that question is "no" as well. But once the questions are firmly separated, it is no longer so clear that it is unacceptable to count the doctor's action as among the causes of the patient's death. (Certainly, the doctor's action was one of the things that helped cause it to be the case that the patient died at 4:12.) At any rate, what *is* clear, given the manifest possibility of confusion, is that it is unwise to use this case to generate substantive conclusions about the metaphysics of causation or of events.

Fifth way: The recalcitrant intuitions conflict with sacrosanct general principles about causation, and so ought to be dismissed.

One such principle very commonly appealed to is, roughly, that like cases be treated alike. Thus, in "Causation as Influence", Lewis has us imagine a case where it is intuitively correct to say that someone's birth is a cause of his death; by appeal to the foregoing principle, he uses this case to argue against the oft-cited complaint that counterfactual analyses have the unacceptably unintuitive consequence that births cause deaths. The example also illustrates how more than one general principle might be brought into play: for one can certainly appeal to transitivity to connect up an agent's birth with his death, as well.

Or again, one might respond to alleged counterexamples to transitivity simply by refusing to give up that principle, perhaps on the basis that its intuitive appeal outweighs the intuitions driving the counterexample, or on the basis that it plays too essential a role in one's analysis. (For example, one might think that only by an appeal to transitivity can an analysis handle certain kinds of pre-emption cases.)

Now, someone who pursues this strategy—insisting that an intuition, however firm, be rejected for the sake of some sacrosanct general principle—will help her case immeasurably if she can add some explanation for why intuition has been so misled. Thus, this strategy is most effectively combined with one or more of the other strategies being discussed here. (As a nice example, consider Beebee's case, in a "Causing and Nothingness" (chapter 11), that there is no causation by omission: She supports this counter-intuitive position in part by appeal to the general principles that causation must relate events, and that omissions cannot be thought of as a species of event, and in part by providing the explanation already discussed for why our intuitions about alleged cases of causation by omission should not be considered trustworthy.) But observe that it may be far less urgent to provide such an explanation if one is not engaged in conceptual analysis: for then one may be able to argue that the given divergence from intuition is the price one must pay for synthesizing a concept that is clean, theoretically attractive, etc.

Sixth way: The unintuitive verdict issued by the theory is not false—it's just odd to say, for pragmatic reasons.

Here for example is Lewis, responding to a worry about his account of causation by omission:

> One reason for an aversion to causation by absences is that if there is any of it at all, there is a lot of it—far more of it than we would normally want to mention. At this very moment, we are being kept alive by an absence of nerve gas in the air we are breathing. The foe of causation by absences owes us an explanation of why we sometimes do say that an absence caused something. The friend of causation by absences owes us an explanation of why we sometimes refuse to say that an absence caused something, even when we have just the right pattern of dependence. I think the friend is much better able to pay his debt than the foe is to pay his. There are ever so many reasons why it might be inappropriate to say something true. It might be irrelevant to the conversation, it might convey a false hint, it might be known already to all concerned…. (Lewis, "Causation as Influence, chapter 3)

We think that care must be taken in executing this gambit. In general, solid reasons should be given why certain judgments should be explained away by appeal to pragmatics, while others should be incorporated into the analysis. (Observe, in this regard, how silly it would be to offer as a theory of causation the following: Every event causes every later event; it is just that in many cases, it is odd to say so, for pragmatic reasons.) What's more, it is one thing to wave one's hands toward some pragmatic account—as Lewis does here—and quite another to provide the specifics

(including: an account of exactly *which* of the many principles governing the pragmatics of communication should be appealed to). We think that unless they are provided—or unless it is fairly clear how they can be provided—the reader deserves to be suspicious of such appeals.

Still, there are other problem cases to which the "pragmatic" treatment seems exactly what is needed. Suzy is in a room with a match. She strikes the match, lighting it. It is perfectly intuitively clear that among the causes of the lighting is the striking. It is much less intuitive—but a consequence of most every analysis of causation, all the same—that among the causes of the lighting is the presence of oxygen next to the match. But we think it a serious mistake to count this a counterexample—at least, without first investigating the possibility that for straightforward pragmatic reasons, reports of causes are typically expected to be reports of *salient* causes. For compare the following scenario: Until recently, all air had been pumped out of the room; only shortly before Suzy entered it with her match was air pumped back in. Now, when asked about the causes of the lighting, we find it quite appropriate to cite the presence of oxygen. Are we to expect this shift in intuitive judgments about what is appropriate to say to reflect a shift in the *truth* of the claim that the presence of oxygen is a cause of the lighting? That strikes us as foolish. Much better to suppose that in altering this scenario, we have altered which features of it are specially salient—and to remember that considerations of salience will weigh heavily on our judgments about what is appropriate to report as a cause.

Seventh way : Intuitions about the case are too easily buffeted about to be taken seriously.

As an example of how such buffeting can work, consider again the case of preemptive prevention in which Billy throws a rock at a window, Suzy blocks it, but Sally was waiting behind Suzy to block the rock, if necessary. Emphasize Sally's idleness, and you can easily secure the intuitive judgment that Suzy prevented the window from breaking. But as Maudlin notes in his contribution ("Causation, Counterfactuals, and the Third Factor", chapter 18), there are other ways to describe such cases. In particular, think of the window and Sally as constituting a single system, remembering that Sally is perfectly able and willing to block the rock. We might then describe this system as a "protected window"—which, because of the state it is in, is *not under any threat* of

damage from the rock. Conceived that way, the case prompts a very different intuitive judgment: Suzy does not prevent the window from breaking, because, after all, the window was never under any threat of being broken.

Now, it seems to us a fascinating phenomenon—worth much more attention than it has been given in the literature—that some cases evoke intuitions that are easily subject to such buffeting, whereas others do not. For example, when Suzy and Billy both throw rocks at the window, and Suzy's gets there first, no amount of redescription of the case will reverse the intuitive judgment that only Suzy's throw is a cause of the shattering. If one is engaged in conceptual analysis, then, this strategy surely backfires: for one thing we should expect from any successful such analysis is an explanation of what distinguishes cases that are subject to buffeting from cases that are not. But if one is engaged in something more like conceptual synthesis, then this strategy might well be appropriate.

Eighth way: The case is too outré for intuitions about it to be of much concern.

As examples, one might want to dismiss from consideration any intuitions about cases that involve action at a temporal distance, or backward causation, or indeterminism at the level of the fundamental laws. Now, we have already seen several reasons why one might set aside such cases: intuitions about them might not be very firm; or there might be reason to suspect that our intuitive judgments cannot easily distinguish such cases from other, less problematic cases; or the judgments about these outré cases might conflict with sacrosanct general principles about causation; etc. But we have in mind something different here, which is that these cases be dismissed *simply* because they are outré in some readily recognizable respect.

To see what this amounts to, observe that the conceptual analyst has no business whatsoever pursuing this strategy. If, for example, we construct a case involving action at a temporal distance—and the case passes all other reasonable methodological tests—then any decent account of how our ordinary concept of causation works must respect the data concerning its application to the given case. On the other hand, someone who constructs an account of causation with the aim of precisifying, reforming, or in some other way altering or replacing our ordinary concept can

perfectly well insist, at the outset, that she only intends her account to cover causation under determinism, or under the assumption that there is no backward causation, or action at a temporal distance, etc. Imagine that her account succeeds brilliantly within its intended domain; and then consider how very foolish it would be to reject it out of hand because it had not been extended to cover a broader domain. We do not mean to suggest that there would be no interest in trying to extend the account; of course there would. But even if there were in principal obstacles to doing so, this would not rob the account of its philosophical interest or utility.

There is, in addition, a second point to make—one which, though modest, is very often overlooked, with unfortunate results. When a philosopher sets out to give an account of causation, and it is clear at the outset that there will be great difficulty in extending the account to cover a certain range of cases (causation under indeterminism, backward causation, etc.), it is all too tempting to simply fix on this limitation as a reason to dismiss the account out of hand. We consider that attitude mistaken, and urge, in opposition to it, a modest methodological pluralism: the topic of causation is difficult enough that it is worth pursuing avenues of investigation even when, at the outset, it seems clear that they will not give us everything we want (i.e., will not give us an account which covers a range of cases that fall outside certain well-defined limitations). A philosopher can therefore reasonably dismiss outré cases from consideration simply because she means—at least for the moment—to limit her ambitions. She is guilty of no philosophical error in doing so. (As an example in this volume, Kvart's contribution (chapter 15) lays out an account of causation under indeterminism that he quite forthrightly observes would be difficult to extend to the deterministic case; it strikes us as no less interesting and important for all that.)

Let us move away now from methodological concerns, and back to properly metaphysical concerns, focusing in the next sections on two questions that confront any account of causation, counterfactual or otherwise: Is causation transitive? And: What exactly are the causal relata?

6. Transitivity

We normally think of the causal relation as a *transitive* relation: if *c* causes *d*, and *d* causes *e*, then *c* causes *e*. The simplest analyses of causation based on sufficiency under laws generate a

transitive causal relation immediately: when *c*, together with the right laws, is sufficient for *d*, and *d*, together with the right laws, is sufficient for *e*, then *c*, together with the right laws, is sufficient for *e*. Analyses of causation based on counterfactual dependence typically need to guarantee transitivity by taking causation to be the ancestral of the dependence relation, as dependence relations (e.g., straightforward counterfactual dependence, or the kind of counterfactual covariation that figures in Lewis's new "influence" account) are typically not transitive.

Some (e.g. Hall: see chapter 7, "Causation and the Price of Transitivity") suggest that the claim that causation is transitive should be treated as a sort of "bedrock datum"; but that strikes us (Hall now included) as hasty. Certainly the claim has a great deal of intuitive appeal; but rather than simply endorse it unquestioningly, we ought to investigate the source of that appeal. Now, while we have nothing decisive to offer on this score, we think it suggestive that the way in which causal claims are often justified seems to *presuppose* transitivity. Someone claims that *c* causes *e*. Why? Because *c* causes *d*, which in turn causes *e*.

This kind of reasoning is commonplace. But if we could not assume that the causal relation were always transitive, then it is not clear why we would ever be entitled to it. It would seem, rather, that in order to determine whether *c* causes *e*, for any *c* and *e* that are not directly linked (i.e., for any *c* and *e* that are causally related, if at all, only *by way of* intermediate events) we would have to determine of every link in the causal chain whether it prevented or allowed for transitivity—whether it was the sort of *c-d-e* link that licensed the inference from "*c* causes *d*" and "*d* causes *e*" to "*c* causes *e*". This would be a serious problem indeed, since (a) it is not clear that we could always distinguish between the types of causal connections that were transitive and the types that were not, and (b) we would, on the face of it, need to have a detailed account of *every* link in the causal chain. Providing such an account would apparently require us to determine the minimal units of the causal chain in question—which, among other things, would involve a fuller specification of the causal relata (usually taken to be events) than theorists have been able to develop.

Hence, it appears that if the causal relation were not transitive, many of our everyday causal claims would be unjustified. (For related discussion, see Lewis, "Causation as Influence", sections 2.2 and 2.3.)

Unfortunately, there are some compelling examples that suggest that causation is *not* transitive. The examples often exhibit the following abstract structure: *c* does something that threatens to prevent *e*. However, *c* also causes *d*, which in turn helps bring about *e* in spite of the threat.[12] For example, a train rushes towards a fork in the tracks. If a switch is flipped, the train will take the left track, if the switch is left in its original position, the train will take the right track. Further on, the left and the right tracks merge, and just after they meet, a damsel in distress is tied to the tracks. If Jill flips the switch, and the train runs over the damsel, should we say that Jill's flip is a cause of the death of the damsel? Obviously not, says intuition. But if the flipping of the switch is a cause of the train taking the left track, and the train running on the left track is a cause of the train's merging back on the main track, and the train's being on the main track is a cause of the death of the damsel—and, finally, if causation is transitive—then Jill's flip is a cause of the death. Notice the abstract structure: the flip threatens to prevent the death—by diverting the train from a track that would have led it to the damsel—but simultaneously does something that helps "undo" the threat—by diverting it *onto* another track leading to the damsel.

A similar kind of example[13] discussed by several of the papers in this collection involves a bomb that is placed in front of someone's door. The bomb is set to explode in 5 minutes. But just before it explodes, a friend comes along and defuses it. As a result, the intended victim continues to live. The presence of the live bomb is a cause of its being defused, and the bomb's being defused is a cause of the intended victim's continued existence. If transitivity holds here, then the presence of the live bomb is a cause of the intended victim's continued existence. Note again the abstract

---

[12] Note that this cannot be the *whole* story: for ordinary cases of preemption such as that exhibited in figure 2 likewise exhibit this structure.

[13] Suggested by Hartry Field, who learned it from Ellery Eells.

structure: the presence of the live bomb threatens to cut short the victim's life, but simultaneously "undoes" this threat by attracting the attention of the friend.

The manifest similarity in structure between these two cases should not make us overlook one difference—a difference that *may* be crucial. In the case of Jill's flip, there is, straightforwardly, a spatiotemporally continuous causal chain running from the flip to the damsel's death; it is "causal" in the sense that each constituent event is, unproblematically, a cause of the immediately subsequent events. The question is only whether being linked by such a sequence of causes is enough for the endpoints to count as causally related. By contrast, in the case of the bomb there is no such spatiotemporally continuous causal chain linking its presence to the continued life of the intended victim. Rather, the case is much more like Hall's "inert neuron network" (figure 11):

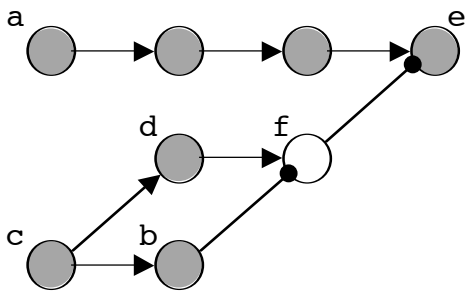

Figure 11

Here **a** fires, stimulating **e** to fire. Meanwhile, **c** fires, stimulating both **b** and **d** to fire. **b**'s firing prevents the signal from **d** from stimulating **f** to fire, thereby safeguarding the firing of **e**—for if **b** had not fired, **e** would not have fired. If such dependence by "double prevention" suffices for causation, then *b* is a cause of *e*. And, obviously, *c* is a cause of *b*. So if, in addition, causation is transitive, then *c* is a cause of *e*: the neuron-world analog of the bomb's presence "causing" the intended victim to continue living. As in the bomb example, no spatiotemporally continuous causal chain connects *c* to *e*. (Unless, perhaps, it is a chain partially constituted by *omissions*. But even if so, that fact does not erase the distinction between this kind of case and the "switching" cases of which Jill's flip is an instance. And see Hall 2002a for doubts that one can always interpolate an appropriate sequence of omissions.)

This distinction might matter, because there is some reason—albeit far from decisive—to hold that dependence by double-prevention does not suffice for causation. If so, then a whole class of alleged counterexamples to transitivity—namely, those that exhibit the structure of figure 11—can be dismissed. Pursuing a "divide and conquer" strategy in defense of transitivity, one could then try to find grounds for dismissing the remaining counterexamples. Hall's "Causation and the Price of Transitivity" takes this approach.

But there are other strategies one can employ to address the problems. David Lewis, in his "Causation As Influence," chapter 3, denies that these cases are counterexamples to transitivity. Lewis argues that as long as the requisite dependence relations hold between links in the chain, transitivity holds, *prima facie* intuitions to the contrary. Lewis traces our mistaken rejection of transitivity to misgivings we might have about accepting preventers as causes, accepting events that initiate deviant or unusual paths to their effects as causes, or to a residual inclination to think that for *c* to cause *e*, *e* must depend counterfactually on *c*.

Stephen Yablo, taking the opposite tack, takes the cases that raise problems for transitivity to be so important as to justify acceptance of a strikingly novel counterfactual account of causation. In his "Advertisement for a Sketch of an Outline of a Proto-Theory of Causation," chapter 5, Yablo defends an analysis of causation that tells us the flip, the bomb or the neuron that initiates the inert neuron network can't count as causes merely because they bring about threats to the effect but simultaneously cancel those threats. There are deeper requirements (such as the effect's depending on the putative cause under conditions that are suitably natural) that an event must meet in order to be a cause.

Cei Maslen, in "The Context Dependence of Causation," chapter 14, argues that causation is a three-place relation between a cause, an effect, and a contrast event: when *c* causes *e*, *c* is a cause of *e* relative to a contrast event *c**. Maslen argues that in cases where contexts (which fix the contrast events) are incompatible, causation is not transitive. She considers several cases that seem to cause problems for transitivity and resolves them under the contrast account.

There is a very different kind of example that needs discussing. Here is an instance (thanks to Judith Thomson for providing it): Suzy lights a firework, which shoots up in the air and explodes in a brilliant shade of red. The light triggers a nearby photodetector that is sensitive to that particular shade of red. Meanwhile, young Zeno, a boy in the crowd, covers his ears in response to the loud explosion. Plausibly, Suzy's lighting the firework counts as a cause both of Zeno's action and of the triggering of the photodetector. But now let us complicate matters: Earlier in the day, Billy added a chemical to the firework that would guarantee that it would explode in the given shade of red; if he hadn't, the explosion would have been yellow. What did Billy's action help cause? Well, the triggering of the photodetector, certainly. But, intuitively, *not* Zeno's action: Billy had nothing to do with that. The problem is this: How are we to secure the first result without undermining the second?

Billy's action does not help bring about the triggering in any mysterious, action-at-a-temporal-distance manner. Rather, there is an ordinary causal chain of events leading from one to the other—by way of the explosion. It therefore seems that in order to count as a cause of the triggering, Billy's action must *also* count as a cause of the *explosion*, since it was only *by way of* the explosion that he helped bring about the triggering. But the explosion also caused Zeno to put his hands over his ears. So why must we not therefore conclude that by adding the chemical, Billy likewise was a cause of Zeno's action?

Looked at a certain way, the case can seem to be another counterexample to transitivity: Billy's action was a cause of the explosion—it had to be, else we could not get it to come out a cause of the triggering of the photodetector. The explosion, in turn, was a cause of Zeno's action. But Billy's action was not a cause of Zeno's action. So perhaps we have, here, yet more reason to give up transitivity.

If so, it is an entirely different kind of reason from that which applied, say, to the inert neuron network: for Billy's action does not both initiate a threat to Zeno's action and simultaneously do something to cancel that threat. Moreover, we cannot in good philosophical conscience stop with a

denial of transitivity, for a thorny puzzle remains: Why is it that *Suzy's* action counts as a cause of Zeno's, by way of the explosion, but Billy's does not?

Some will be impatient for what they consider the obvious response: there is not just *one* explosion, but at least *two*—distinguished by the fact that it is somehow constitutive or essential to one of them, but not the other, that it be an explosion in the given shade of red. Call these explosions "*e*" and "*e*+", the latter being the one that is constitutively or essentially red in the given shade. Then we have what seems a transitivity-preserving solution: Billy's action causes *e*+ but not *e*; Suzy's action causes both *e* and *e*+; *e*+ causes the triggering but not Zeno's action.

An attractive story, perhaps, although many will find the ontological inflation of events *un*attractive. We should not accept it without taking several steps back, and considering with more care than we have so far the question of what, exactly, the causal relation *relates*. That is what L. A. Paul does in "Aspect Causation," chapter 8. She discusses a case[14] that appears to create problems for transitivity, and argues that this appearance is illusory—closer inspection shows that in such cases, transitivity fails to apply. Causation may well be transitive; it is, rather, an underlying tension between extant theories of events coupled with reductive analyses of causation that gives rise to these seeming counterexamples. Here is the example that Paul considers: Suzy breaks her right wrist, which causes her to write her philosophy paper with her left hand. The paper is then sent off to a journal, which accepts it. If the breaking causes the event of the left-handed paper-writing, and the paper writing causes the acceptance, and if causation is transitive, it seems—contrary to intuition—that the breaking caused the acceptance.

Paul shows that none of the major contenders for a theory of events coupled with a theory of causation succeed against such examples, and argues that this makes trouble for theories of event causation. She exploits this trouble in order to argue in favor of taking *property instances* (what she calls "aspects") rather than events to be the causal relata. (We discuss related issues involving the nature of the causal relata below.) Observe how this maneuver will handle the foregoing case of the

---

[14] The case appears in McDermott 1995a.

loud, colorful explosion: Billy's action causes the explosion to have a certain aspect—namely, to be an explosion in that particular shade of red. This aspect in turn is responsible for the triggering—but not for Zeno's covering his ears. Transitivity is preserved. Still, as Hall points out in his "Causation and the Price of Transitivity," chapter 7, making room for aspects is not enough to resolve what he calls the "hard cases" that threaten transitivity, so even if a different account of the causal relata helps, serious problems remain. We will nevertheless set them aside, and focus squarely on the question Paul's paper brings to the fore: What does causation relate?

7. The causal relata

Most contemporary analyses of causation hold that causation is a relation between *events*, where these are taken to be particulars, entities that occupy spatiotemporal regions. But unless we are prepared to take events as unanalyzable primitives, to have an acceptable (reductive) analysis of the causal relation as a relation between events requires an acceptable theory of eventhood. The theory of events thus counts as a subtheory of a complete theory of causation.

By wide agreement, such a theory will broaden the category of event to include items never ordinarily so-called. When Suzy throws a rock at the window, breaking it, we naturally tend to think that there is just *one* sequence of events—the one initiated by Suzy's throw—converging on the effect. But it is far better, at least for the purposes of systematic metaphysics, to see this effect as standing at the intersection of *two* sequences of events: there is the interesting sequence just mentioned, and then there is the quite boring sequence consisting in the continued presence of the window, up to the moment it shatters. More generally, a proper theory of events almost certainly must count as such things that we would ordinarily classify as states, or standing or background conditions.

Next, there is widespread agreement that a theory of events must have the resources to tell us when two events, though non-identical, fail to be wholly distinct—e.g., because one is part of the other, or they have a common part, or they stand in some sort of intimate logical relationship. For, as we saw in section 4, a theory unable to draw such distinctions will almost certainly leave the

account of causation that makes use of it easy prey for various counterexamples, in the form of not-wholly-distinct events that stand in relations of counterfactual dependence.

But within these broad parameters, there is no consensus as to which theory of eventhood is the correct one. This lack of consensus traces in part to conflicting intuitions—in particular, about when we have just one event, and when (and why) we have more than one. You walk to class. As you walk, you talk to your companion. Is your walking one event and your talking a numerically distinct event? Or are they the same event? There are considerations on both sides. On the one hand, the walking and the talking occurred at the same time and were performed by the same person, so common sense might dictate that they are the same event. On the other hand, the walking and the talking seem to stand in different causal relations to subsequent events: your walking made your legs tired, while your talking gave you a sore throat.

In general, we can characterize the individuation conditions put forward in different theories of events along a continuum from *coarse grained* to *fine grained*. Roughly, the more fine grained a theory makes event identity, the more events exist, according to the theory. So a theory that held that the walking and talking were one and the same would be more coarse grained than a theory that distinguished them; this theory in turn would, if it recognized but *one* talking, be more coarse grained than a theory that distinguished your talking *loudly* from your talking *simpliciter*.

It's one thing to insist on such distinctions; it's another to explain them. Among those who want to draw them, probably the most popular explanation appeals to differences in the possible conditions of occurrence—or "essences"—of events.[15] You could have walked without talking, and vice versa; so your walking and talking are distinct events. More ambitiously, perhaps your talking *loudly* is a different event from your *talking*, because the latter, but not the former, would have occurred if you had been talking softly. If so, then the two events are alike in all categorical respects: not only do they occupy the same region of spacetime, but they are both loud (one essentially, one accidentally so), both energetic, etc. Notice that once the door is open to events that

---

[15] See Lewis 1986d. For a closely related essentialist theory, see Yablo 1992a, 1992b, and 2000.

differ *only* along such modal dimensions, it will be difficult to close it; that is, the number of events that one must recognize as being perfectly categorically coincident will probably be enormous. There is the talking, the talking that is essentially not quiet, the talking that is essentially loud, the talking that is essentially exactly *that* loud, etc.—and we have been drawing only crude distinctions, about only one of the many aspects of these events. This ontological proliferation leads to trouble, as we will shortly see.

Theories of events that do not draw distinctions in this way—either because they draw them in some other way[16], or not at all—still face an important question concerning essences. Consider your walking: what are the possible conditions under which *it*—and not just some event very much like it—could have occurred? To the extent that these conditions are very restrictive, your walking is modally *fragile*—unable, as it were, to survive in all but a few alternate possible worlds.

It is important to recognize that the question of how fragile events are cross-cuts the question of how many there are. For example, one could hold a highly coarse-grained theory—say, one according to which there is at most one event per region of spacetime—without committing oneself as to how modally fragile these events are. Or one could hold a highly fine-grained theory while remaining uncommitted. Of course, the questions *can* intersect: the account just considered, that distinguishes a multitude of perfectly coincident events by their differing conditions of occurrence, is highly fine-grained in a way that automatically commits it to the existence of plenty of events that are extremely modally fragile—and, for that matter, plenty of events that are not fragile at all, and plenty of events that are somewhere in between. (For example, your talking *loudly* is more fragile than your talking *simpliciter*, etc.)

Among theories that hold that (all) events are fragile, towards one extreme is a view according to which *any* difference in time, place, or manner of occurrence would make for a numerically different

---

[16] See for example Kim 1973a and 1980. According to Kim, each event corresponds to a triple of a "constitutive" time, particular, and property. It is not entirely clear what it means for such elements to be constitutive. But Kim explicitly distances himself from the view that they are constitutive in the sense that it is essential to the event's occurrence that the given particular has the given property at the given time.

event.[17] Why adopt such a position? We can recognize one important reason—and also an unfortunate consequence—if we return to late preemption. Recall our paradigm example: Billy and Suzy throw rocks at a window, with Suzy's reaching it first, breaking it. How can we secure the result that Suzy's throw, and not Billy's, causes the breaking? Perhaps by insisting that the actual breaking is quite fragile: had the window broken a few moments later, that would have been a numerically different breaking. Then, arguably, the breaking that actually occurs counterfactually depends on Suzy's throw, but not on Billy's.

Many people find this view compelling until they realize that tightening the individuation conditions for events in this way results in a landslide of causal connections, too many of them unwanted. (It should also be remembered that variations on late preemption yield examples immune from this treatment.) Your sore throat occurred at precisely 4 p.m., but if you had not stopped by your office to pick up your bag before you commenced your walk (and thus your discussion), you would have had a sore throat at 3:59 p.m. Under the fragility account, that sore throat would be a *numerically distinct event* from the sore throat at 4 p.m. Hence, standard regularity and counterfactual accounts of causation will count picking up your bag from your office a cause of the sore throat that actually occurred.

Although defenders of fragility accounts recognize that many more events will count as causes and effects than untutored intuition will be comfortable with, they believe the cost is worth the benefit. David Coady, in "Preempting Preemption," (chapter 13) argues that a counterfactual analysis of causation employing fragile events will solve a number of problems, and along the way gives an able defense of the fragility strategy.[18]

Notice that the advantages that accrue to this strategy do not obviously extend to an account that, while recognizing the existence of a highly fragile window-breaking, also recognizes less fragile

---

[17] *Towards* one extreme, but not all the way to the end: One could add, for example, that any difference in causal origins would make for a different event.

[18] Lombard 1986 also holds that events are fragile with respect to time, so could not have occurred at any time other than the time they actually did.

(and perfectly coincident) breakings, all differing in their essences. On such an account, there is a window-breaking that necessarily occurs just when it does; perhaps that breaking counterfactually depends on Suzy's throw alone. But there is also a breaking that could have occurred as much as a few moments later. What caused *it*? It's not even intuitively clear what the answer is, but at any rate it would be an embarrassment for an account of causation if it held that *nothing* caused this event (e.g., because no event could be found on which it depends, etc.). What reason could there be for admitting such uncaused events into one's ontology?

In fact, the reason standardly given is exactly that we need such events as effects (and causes). More precisely, perfectly coincident events differing only in their essences are supposed to help us draw certain causal distinctions. For example, suppose Billy's throw makes a slight difference to the way in which the window shatters, perhaps by way of the gravitational pull his rock exerts on Suzy's. We would like to say: Billy's throw doesn't cause the breaking, but it *does* help cause that breaking to happen in just the way it does. Or, returning to the example considered toward the end of last section, Billy's action does not cause the firework's explosion, but *does* cause it to be in the color it is. Distinguishing a breaking that necessarily happens in just the way it does provides us with an event to be an effect of Billy's throw; distinguishing an explosion that is essentially red provides us with an event to be an effect of Billy's earlier action (and, by the way, to be a *cause* of the subsequent triggering of the photo-detector).

Of course, this reason for drawing purely modal distinctions between events provides no ingredients for solving the problem posed by late preemption; the theorist needs to turn elsewhere for such a solution. There is a worse problem. Consider a very simple situation, in which Suzy, all alone, throws a rock at a window, breaking it. The following sentence is true—and *determinately* so: "Suzy's throw is a cause of the breaking." But to what do the expressions "Suzy's throw" and "the breaking" refer? There are far too many eligible candidates for it to be at all plausible that their reference is *determinate*. So what? Why not just recognize a (large) number of admissible assignments of referents to each expression? Well, because doing so leaves it obscure how our sentence gets to have a determinate truth value. Suppose we assign a comparatively fragile event to

"Suzy's throw" and a comparatively robust one to "the breaking"; then our effect will in all likelihood not depend on our cause. Nor is this trouble just for a counterfactual analysis: reverse matters, assigning a fragile event to "the breaking" and a robust one to "Suzy's throw", and a typical regularity account will have trouble making the latter out to be a cause of the former. The upshot is that on standard accounts of causation, there will be admissible assignments of referents that make our sentence *false*. And so even if there are admissible assignments that make it *true*, the sentence will not come out—as it should—*determinately* true.[19]

Why not opt for a more parsimonious account of events—holding, for example, that events are individuated by the spatiotemporal regions they occupy? But such a seemingly attractive theory itself has unwelcome consequences. If we had a metal sphere that as it rotated also heated up, we would have to identify the event of the rotating of the sphere with the event of the heating up of the sphere—they are the same event because they occur in the same spatiotemporal region. But our earlier reasons for wanting occasionally to distinguish coincident events remain in full force, for it seems easy to imagine a situation in which the sphere's rotation has different effects from its heating.

A different approach to causal relata bypasses at least some of these problems by holding that the causal relata are not events after all. (Or at least, the causal relation can only be said to hold between events in virtue of holding between something else.) Various alternative accounts of the causal relata have been put forward, the most plausible of which take them to be property instances, facts, states of affairs, or propositions.

Those who hold that the causal relata are property instances often argue that causation obtains between instantiations of universals or between tropes. A different but related view holds that causation obtains between states of affairs, where states of affairs consist in particulars having properties. Such views can still accept that events are causal relata in a derivative sense, since events may well be constructed from property instances or states of affairs.

---

[19] See Bennett 1988 for an expert statement of this problem.

In "Aspect Causation," chapter 8, Paul argues that we should take the causal relata to be property instances (*aspects*), as this can help resolve some outstanding problems with transitivity and allows us to bypass the need to provide individuation conditions for events. If the causal relata are property instances, then we can construct a reductive analysis of causation independently of an analysis of events—a result that, given the lack of consensus over what counts as a satisfactory theory of events, philosophers of causation should welcome.

There is a quite different motivation for abandoning or at least qualifying the default position that events are the primary causal relata. It stems from the need to accommodate causation involving *omissions*: prevention, causation by omission, and (arguably) causation by double prevention. (See the contributions by Lewis, Beebee, Menzies, Mellor, and Armstrong.) Focus just on causation by omission, as when Billy's failure to show up for their lunch date causes Suzy to become disappointed. The possibility of this kind of causation creates an immediate difficulty for those who would hold that causation must relate events. For on the face of it absences (omissions) are not particulars at all, and so not events; what are we to make, then, of causation "involving" them? We seem to find ourselves with a causal relation that is missing a relatum. D. H. Mellor, in his "For Facts as Causes and Effects," chapter 12, responds by defending the view that *most* causes and effects are not particulars, but rather *facts*. Mellor takes facts to be states of affairs expressed by true sentences, statements or propositions.[20] An advantage of Mellor's position is that it allows him to handle cases involving negative causation quite naturally. It is an unhappy philosopher who is forced to believe in negative *events*, but Mellor by contrast can quite serenely grant that there are negative *facts*.

Lewis's response, in "Causation as Influence," chapter 3, and especially "Void and Object", chapter 10 is to insist that when there *are* causal relata, they are invariably events—but to go on to assert that in cases of causation by omission or prevention, one of the relata is missing: "So I have to say that when an absence is a cause or an effect, there is strictly speaking nothing at all that is a

cause or effect. Sometimes causation is not a relation, because a relation needs relata and sometimes the causal relata go missing." ("Causation as Influence", p. XXX) Nevertheless, he claims, in all such cases there will be an appropriate pattern of counterfactual dependence—not between two events, but, e.g., between an event and the fact that no event of some specified type occurs. (Thus, Suzy's sadness counterfactually depends on the fact that no event of the Billy-showing-up type occurs.) Why this relation of dependence does not qualify the foregoing fact as a *cause* is, to our minds, left a bit unclear.

Beebee responds to problems involving absences by rejecting causation by omission. For Beebee, causation always relates particulars; hence, there can be no causation involving absences. She argues that commonsense intuitions about causation by omission make trouble both for those who defend causation by omission and those who reject it. For example, she must deny that Billy's absence causes Suzy's disappointment. But those who disagree typically hold that mere counterfactual dependence suffices for causation by omission; consequently, they must hold that the failure of Bill Gates to show up with a fat check for Suzy also causes her disappointment. Beebee accounts for the central role that causation by omission seems to play in our ordinary causal claims by arguing that common sense fails to discriminate between causation and explanation.

Menzies takes a new approach to the problem of how our intuitions about omissions seem to generate conflicting results, arguing that we can develop an account of causation based on relevant difference-making that distinguishes between absences as causes and absences as "mere conditions." (The distinction between causes and conditions is relative to a field of normal conditions generated by a causal model.) Menzies' account is constructed so as to avoid the sorts of spurious causes that Beebee cites, and that make trouble for a non-discriminating counterfactual account of causation by omission.

Recognizing the special challenges that causation involving absences presents suggests that it might be helpful to distinguish different kinds of causation. There is garden variety causation, as

---

[20] Others have also defended the idea that the causal relata are facts; e.g., Jonathan Bennett gives a well argued

when Suzy's throw causes the window to break. There is causation by omission, as when Billy's failure to stop her likewise causes the breaking. There is prevention, as when (this time) Billy stops the rock mid-flight, preserving the window. There is double prevention, as when Sally knocks Billy aside before he can block the rock, thus contributing to the window's breaking. And perhaps there are still more exotic varieties of omission-involving causation.

It may well be that the different kinds of causation require different explications. (Both Hall, in "Two Concepts of Causation", chapter 9, and Armstrong, in "Going Through the Open Door Again", chapter 19, argue as much.) If so, then it comes as no surprise that uniform analyses of causation are vulnerable to counterexamples and inadequate for complicated examples where different kinds of causation are combined. Here it is well to recall some of the problems involving transitivity discussed in the last section, where omissions played a central role in the structure of the case (the inert network is a good example). Hall in particular argues that such cases threaten transitivity only if we hold that transitivity must apply to the kinds of causation—in particular, causation by double prevention—featured in them.

The problem of explaining causation by and of omission, and hence by extension explaining what it is to prevent events from occurring or to allow events to occur, is an area of continuing research. The importance of the problem is highlighted by the special roles that preventing and allowing seem to play in related areas of inquiry; for example, accounts of our ascriptions of moral and legal responsibility.

8. Summaries of the contributions

Here we provide capsule summaries of the contributions that follow.

Chapter 2: Jonathan Schaffer, "Trumping Preemption"

---

defense of the view that causes and effects are facts in his 1988.

Schaffer develops a novel kind of example that poses trouble for most existing accounts of causation. So-called *trumping preemption* occurs when two (or more) processes, each sufficient to bring about some effect, go to completion. But unlike cases of symmetric overdetermination, there is an asymmetry: one of the processes is such that variation in it would have been followed by corresponding variation in the effect, whereas the same is not true of the other. For short, the first process *trumps* the second. As an example, Schaffer imagines a sergeant and a major simultaneously shouting orders to their troops. In cases of conflict, the troops obey the major; but this time the order is the same. Schaffer argues, first, that in cases like this the "trumping" event (e.g., the major's shout) is the sole cause of the effect (the troops' response), and, second, that existing counterfactual accounts of causation cannot accommodate this fact.

Chapter 3: David Lewis, "Causation As Influence"

Lewis replaces the old idea that causation is, at heart, counterfactual dependence between events, with the thesis that causation is, at heart, counterfactual *covariation* between events. He first defines a relation of "influence"—which is, very roughly, the relation an event *c* bears to an event *e* just in case the manner of occurrence of *e* counterfactually depends in a suitably systematic way on the manner of occurrence of *c*. Lewis then takes causation to be the ancestral of influence. As an illustration, Lewis's solution to the problem of late pre-emption—as exhibited, say, in the example of Suzy, Billy, and the broken window—is that Suzy's throw exerts much more influence on the shattering of the window than does Billy's: counterfactually varying Suzy's throw in any of a number of ways will result in corresponding changes in the shattering, whereas the same is not true of Billy's throw—or at least, not nearly to the same extent. For that reason, Suzy's throw, and hers alone, counts as a cause of the shattering. En route to developing this account, Lewis treats a number of central methodological and metaphysical issues, discussing transitivity, causation involving omissions, the varieties of pre-emption, and the reasons for pursuing an analysis of causation at all.

Chapter 4: John Collins, "Preemptive Prevention"

Chapter 5: Steve Yablo, "Advertisement for a Sketch of an Outline of a Proto-Theory of Causation"

Yablo argues that we should return to the original, spartan idea that causation is counterfactual dependence—but understand that dependence to be something he calls "de facto dependence". The idea is, roughly, that *e* de facto depends on *c* just in case, had *c* not happened—and had suitably chosen other factors been held fixed—then *e* would not have happened. As an illustration, Yablo would say that the shattering of the window de facto depends on Suzy's throw because, had Suzy not thrown—and had it still been the case that Billy's rock never struck the window—then the window would not have broken. Obviously, the success of this proposal depends heavily on being able to pick out, in a principled way, which factors should be held fixed; much of Yablo's paper is devoted to this very question.

Chapter 6: Peter Menzies, "Difference-Making in Context"

Menzies argues that Lewis's counterfactual analysis of causation is insensitive to the different context-relative ways in which commonsense distinguishes between causes and background conditions. Menzies develops the conception of a cause as something that makes a difference with respect to an assumed background or causal field by giving a detailed analysis of an account of difference-making in terms of context sensitive counterfactuals.

Chapter 7: Ned Hall, "Causation and the Price of Transitivity"

Chapter 8: L. A. Paul, "Aspect Causation"

Paul shows that reductive analyses of causation that take events to be the causal relata face counterintuitive consequences with respect to the transitivity of causation, and proposes an analysis of causation where aspects, or property instances, are the causal relata. Aspect causation combines

elements of regularity theories with influence (patterns of counterfactual covariation) and takes property instances to be the causal relata. Paul argues that changing the causal relata to property instances resolves some outstanding problems with transitivity and allows the development of a theory of causation that is not held hostage to a theory of event individuation.

Chapter 9: Ned Hall, "Two Concepts of Causation"

Chapter 10: David Lewis, "Void and Object"

Chapter 11: Helen Beebee, "Causing and Nothingness"

Chapter 12: Hugh Mellor, "For Facts as Causes and Effects"

Mellor argues that facts, rather than particulars such as events, are causes and effects. For Mellor, singular causal claims are best rendered in the form "E because C", where C and E are sentences and "because" is a sentential connective; "E because C" is interpreted as equivalent to "the fact that C causes the fact that E." Mellor argues that there is no reason to take causation to be a relation, much less a relation between events, and that this perspective enables negative facts to be causes and effects. For Mellor, facts are superior to events as causes and effects because fact causation can handle causation of or by negative facts much more easily than event causation can handle causation of or by omission.

Chapter 13: David Coady, "Preempting Preemption"

Coady defends what he calls the "naïve counterfactual analysis of causation", that *c* causes *e* iff, if *c* had not occurred, *e* would not have occurred. This analysis is widely believed to be refuted by the possibility of pre-emption, which is a species of redundant causation. Through a careful consideration of the different kinds of events related in paradigmatic examples, Coady argues that

there are in fact no cases of preemption: all such cases are either cases of symmetrical overdetermination or are not cases of redundant causation after all.

Chapter 14: Cei Maslen, "The Context-Dependence of Causation"

Maslen develops an account according to which events cause others event only relative to contrast situations. For Maslen, causation in the world is an objective, *three*-place relation between causes, contrasts and effects, and the truth and meaning of causal statements depends upon the context in which they occur (for it is this context that helps determine the relevant contrast situation). Maslen also argues that her contrast theory of causation makes the intransitivity of the causal relation intuitively plausible and gives compelling analyses of several counterexamples to the transitivity of (binary) causation.

Chapter 15: Igal Kvart, "Causation: Probabilistic and Counterfactual Analyses"

Chapter 16: Murali Ramachandran, "A Counterfactual Analysis of Indeterministic Causation"

Chapter 17: Christopher Hitchcock, "Do All and Only Causes Raise the Probabilities of Effects?"

Chapter 18: Tim Maudlin, "Causation, Counterfactuals, and the Third Factor"

Chapter 19: David Armstrong, "Going Through the Open Door Again: Counterfactual vs. Singularist Theories of Causation"

Armstrong criticizes the counterfactual analysis put forward by David Lewis and defends a singularist theory. Armstrong's singularist theory holds that causation is a two-term relation—i.e., a universal—holding between cause and effect (where causes and effects are states of affairs). For Armstrong, the concept of causation is conceptually primitive. He cites several reasons why we can take ourselves to have observational access to singular causal facts: ordinary language, experiences

such as the perception of pressure on one's body, our awareness of the operation of our will, and psychological experimental evidence. Armstrong also discusses the treatments he favors for several issues concerning causation, including causation by omission, action at a distance, probabilistic causation and the connection between causation and laws.