5

Causation and Pre-emption

Ned Hall and L. A. Paul

1 Introduction

Causation is a deeply intuitive and familiar relation, gripped powerfully by common sense. Or so it seems. As is typical in philosophy, however, that deep intuitive familiarity has not led to any philosophical account of causation that is at once clean, precise, and widely agreed upon. Not for lack of trying: the last thirty years or so have seen dozens of attempts to provide such an account, and the pace of development is, if anything, accelerating. (See Collins *et al.* [2003*a*] for a comprehensive sampling of the latest work.)

It is safe to say that none has yet succeeded. It is also safe to say that the effort put into their development has yielded a wealth of insights into causation. And it is, arguably, from the study of *causal pre-emption*—cases that feature multiple competing candidates for the title of 'cause' of some given effect—that the greatest such wealth has flowed. These cases come in a number of varieties: so-called early and late pre-emption, symmetric overdetermination, and trumping pre-emption. Collectively, they place extremely severe constraints on any philosophical account of causation that can successfully handle them. One of the lessons they have to teach, then, is a lesson about the form that a successful analysis of causation must have.

There is a deeper lesson, a lesson about the nature of causation itself—or, if you like, about the workings of our causal concept. It emerges from close study of the struggles that extant accounts face in trying to provide even remotely attractive treatments of causal pre-emption. It is this: there appears to be a significant and perhaps intractable tension between one strand in our thinking about causation—a strand that emphasizes the need for causes to be *connected* to their effects via intervening processes with the right intrinsic character—and another—one that emphasizes the claim that effects in some sense *depend* on their causes. By the end of this essay this tension will be vividly apparent. One response to it is to insist that a proper account of causation must simply be developed with extreme care, so as to respect both strands in our thinking. Another response is to insist that we must recognize distinct *kinds* of causation, characterized by different fundamental features. And there is always the most pessimistic response, which is to deny that any informative philosophical analysis of causation is possible.

We aim not to settle such questions, but to illuminate them, by providing as accurate a guide as possible to the philosophical terrain inhabited by cases of redundant causation, and by the analyses that have sprouted up around them. To do so we must, out of necessity, make some highly discriminatory choices up front, lest our 'guide' itself require a guide. The next section outlines these choices, and takes care of other preliminaries.

2 Preliminaries

We assume that the fundamental causal relata are *events*. (For alternative views, see Bennett [1988], Paul [2000], and Mellor [2003].) We also assume a broadly reductionist outlook, according to which facts about which events cause which other events are fixed, somehow, by (i) the facts about what happens, and (ii) the facts about the fundamental laws that govern what happens. (For opposing views, see e.g. Anscombe [1971], Tooley [1990], and Cartwright [1983], [1999].) We shall not investigate the metaphysical nature of these laws (but see Maudlin [2003]), although we *will* assume, purely for the sake of simplicity, that they are deterministic, and that they permit neither backwards causation nor causation across a temporal gap.

We pass by a number of issues central to the metaphysics of causation, resting content with pointers to the relevant literature. Thus, we will not consider what form a philosophical account of events should take if it is to serve the purposes of an account of causation; for a sampling, see Davidson [1969], [1970], Kim [1973], Lewis [1986c], Lombard [1986], Bennett [1988], and Yablo [1992]. We will not consider indeterministic causation; but see Lewis [1986*b*], Eells [1991], Menzies [1996], Schaffer [2000*a*], Hitchcock [2003], Kvart [2003], and Ramachandran [2003] for sophisticated treatments. We ignore the possibility of simultaneous or backwards causation, or causation across temporal gaps; we likewise ignore the related problem of how to account for the *asymmetry* of causation, if it is left open that causes need not precede their effects (so that we cannot simply piggyback causal asymmetry on temporal asymmetry); but see Collins *et al.* [2003*b*] for discussion.

We regret that we have not the space to investigate a number of important and interesting cases that do not involve pre-emption, most significantly cases involving *omissions*, and cases that appear to challenge the claim that causation is *transitive*. And certain cases that *do* fall under our topic must also be ignored—cases of trumping (Schaffer [2000*b*]) and symmetric overdetermination in particular. Note however that, although they have attracted a small flurry of interest in the recent literature, cases of trumping turn out on inspection to be nothing more than either cases of symmetric overdetermination in disguise or cases of late pre-emption in disguise; either way, they have nothing new to teach us. Cases of symmetric overdetermination matter more, since

regularity accounts can accommodate them much more easily than can counterfactual accounts. (See Hall and Paul [2002] for comprehensive discussion of all of these topics, as well as even more detailed discussion of early and late preemption.)

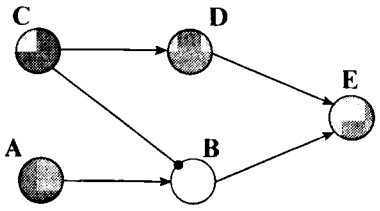

Figure 1

In laying out our examples, we make extensive use of 'neuron diagrams' (popularized by Lewis: see in particular his [1986*b*]). Figure 1 provides a sample. Circles represent neurons, arrows stimulatory connections between neurons, lines with blobs on the end inhibitory connections. A shaded circle indicates that the neuron fires. The order of events is left to right. Thus, in the figure, neurons **A** and **C** fire simultaneously; **C** sends a stimulatory signal to **D**, causing it to fire, while **A** sends a stimulatory signal to **B**. (Throughout this essay, bold capitals name neurons, italicized capitals, events of their firing.) But, since **C** also sends an inhibitory signal to **B**, **B** does not fire. Finally, **D** sends a stimulatory signal to **E**, causing it to fire. Figure 2 shows what would have happened if **C** had not fired:

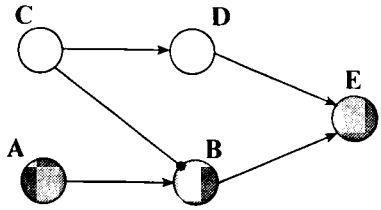

Figure 2

Properly used, neuron diagrams can represent a complex situation clearly and forcefully, allowing the reader to take in its central causal characteristics at a glance. They can also mislead, and their prominence in this essay should not suggest that we think that every interesting and important feature of an example can be 'boiled down' to such a diagram. Where necessary, we will take pains to highlight those features that cannot.

We will focus our attention solely on *regularity* and *counterfactual* analyses of causation, partly because we consider them the most interesting and promising approaches to causation, partly because it is their struggles with redundant

causation that best bring out the lessons we wish to address, and partly, of course, for the sake of brevity. For discussion of other approaches—in particular, 'transference' accounts of the kind found in Fair [1979], Salmon [1994], Ehring [1997], and Dowe [2000]—see Hall and Paul [2002].

3 Counterfactual accounts

Counterfactual accounts of causation begin with the idea that, when E counterfactually depends on C (for short, just 'depends')—that is, when it is the case that, if C had not occurred, E would not have occurred—then C must be a cause of E. It is crucial that the counterfactual here be understood in a 'non-backtracking' sense (Lewis [1973]). If, for example, we ask what would have happened in Figure 1 if **D** had not fired, we do *not* work backwards, considering what, given the laws, would have to have been the case in order for **D** not to fire—for then we will reason that it would have to have been the case that **C** did not fire, hence that **B** *did* fire, hence that **E** fired. Rather, we hold fixed factors contemporaneous with the event we are 'excising'—in particular, the non-firing of **B**—and evolve the resulting state (one in which *neither* **D** nor **B** fire) forward in accordance with the laws. Result: if **D** had not fired, **E** would not have fired. It is a good question, but one we must pass over, whether adequate *non-causal* truth conditions can be supplied for such 'non-backtracking' counterfactuals; see Lewis [1979] for an influential proposal, and Elga [2000] for an important recent critique.

Dependence cannot be a necessary condition on causation, since C can cause E, even though backup processes would have brought about E in C's absence (as Figure 1 already shows, and as we will see in more detail in the sections ahead). But as a sufficient condition on causation, it has struck many philosophers as exactly right—and therefore as an excellent starting point for a full-blown analysis of causation. Scan the literature on causation, and you will find a profusion of such analyses, departing in myriad different directions from this leading idea. We will sketch here what seem to us the three most interesting avenues. (Others will be mentioned in the sections to follow.)

Probably the simplest and most elegant approach comes from Lewis [1973], who analyzes causation as the ancestral of counterfactual dependence: C causes E just in case there is a (possibly empty) set of events $\{D_1, D_2, \ldots, D_n\}$ such that D_1 depends on C, D_2 depends on $D_1, \ldots$, and E depends on D_n.

The second account comes from Lewis's more recent work [2000], [2003], where he replaces the simple relation of dependence with a more complicated relation of what we will call *counterfactual covariation*. Very roughly, E counterfactually covaries with C just in case (and to the extent that) variation in the manner of C's occurrence would be followed by corresponding variation in the manner of E's occurrence. The situation in which C's absence would be

followed by *E*'s absence can be seen as a kind of limiting case. Following Lewis, say that *C influences E* just in case *E* counterfactually covaries with *C* to a sufficient extent. (Also following Lewis, we will leave it vague what counts as 'sufficient'.) Lewis's proposal is that causation is the ancestral of influence.

The third approach has been recently championed by Yablo [2003], and by advocates of the so-called 'structural equations' approach to causation (Hitchcock [2001], Pearl [2000]). Its key idea is to identify causation with what Yablo has called 'de facto dependence': *E* de facto depends on *C* just in case, had *C* not occurred, *and* had other suitably chosen factors been held fixed, *E* would not have occurred. The rider 'suitably chosen' is indispensable, since without it we trivialize the account by letting the factors to be held fixed be the fact that either both or neither of *C* and *E* occur. In addition, the account needs supplementing by clear and systematic truth conditions for this more complex kind of counterfactual. (One doesn't always find such truth conditions supplied; Yablo [2003], for example, completely overlooks them.)

It will help to have an example. Given the popularity that structural equations approaches in particular seem currently to enjoy, we will draw ours from that literature (specifically, from Hitchcock [2001]). Laying out the example requires a brief digression in order to explain what is distinctive about the structural equations approach. According to its advocates, in order to analyze the causal structure of any situation, we must first provide a 'causal model' for it. The elements of this model consist of *variables*; a range of possible *values* for each of the variables; a specification, for each variable, of which other variables (if any) it *immediately functionally depends on*; and, finally, 'structural equations' that describe this dependence.

It is an excellent question what principles should guide the construction of a causal model for a given situation—and in particular whether their application simply presupposes knowledge of that situation's causal structure. Never mind. For neuron diagrams, at least, the task is easy: first, assign a variable to each neuron, which can take on a range of values corresponding to each different way that that neuron can fire, reserving one more value for the situation in which it does not fire at all. Next, stipulate that each such variable immediately functionally depend on the variables for those neurons that have a direct 'incoming' connection to it, either stimulatory or inhibitory. And finally, write the functional equations down so as to capture exactly how the various possible firing patterns for the input neurons to a given neuron will determine whether and how it fires.

The power of causal models resides in their ability to confer systematic truth conditions on counterfactuals whose antecedents specify values for arbitrarily many variables. Consider Figure 1, and suppose we wish to evaluate what would have happened if **C** had not fired, *and* if **B** also had failed to fire. When, as here, the antecedent stipulates the value for some 'endogenous' variable (i.e.

a variable whose value functionally depends on other variables explicitly represented in the model, as the B-variable does here), then in constructing the counterfactual situation we simply *ignore* those functional equations that would otherwise have fixed the value of this variable. Thus, we set the value of the C-variable to 0, of the A-variable to 1 (its actual value), and of the B-variable to 0. We then calculate the values for the D- and E-variables according to the appropriate functional equations, with the result that the D-variable has the value 0 and the E-variable also the value 0. In words, if **C** had not fired, and **B** had (still) not fired, then **E** would not have fired.

Observe that this counterfactual allows us to say that, in a sense, *E does* depend on *C*; for *in fact* **B** does not fire, and if we hold this fact fixed then *E* depends on *C*. More generally, suppose that we have two events, *C* and *E*, and associated variables. And suppose that our causal model of the situation in which *C* and *E* occur describes a path from the C-variable to the E-variable via a sequence of intervening variables, connected by relations of immediate functional dependence. Lastly, suppose that there are one or more variables that are *not* on this path, such that *E* depends on *C given* that they are held fixed at their actual values. Then, adopting Hitchcock's terminology, we can say that the given path from *C* to *E* is an 'active route'. A simple proposal results: *C* is a cause of *E* iff there is an active route from *C* to *E*. For example, in Figure 1 the **C**–**D**–**E** route is active, as witness the fact that, if **C** had not fired and **B** had also not fired, then **E** would not have fired. By contrast, there appears to be no active route from *A* to *E*: for the only candidate is the **A**–**B**–**E** route, and holding fixed any combination of the C- and D-variables fails to make it the case that *E* depends on *A*.

We emphasize that this is not the only way to construct a de facto dependence approach, or even a structural equations variant thereof. But it will provide an attractively simple illustration of the approach in the pages ahead; observe in this regard that it is crystal clear what constrains the choice of factors to be 'held fixed'. Readers are invited to contrast Yablo's [2003] discussion of this matter, which is vastly more complicated.

4 Regularity accounts

What have traditionally been called 'regularity' accounts of causation have been guided by two quite distinct ideas. The first is that causal relations between events should be analyzed as *instances of lawful regularities* (Davidson [1967]). We think that it carries little promise, and will focus on a second idea, which is that what is distinctive of a cause *C* of some event *E* is that *C suffices* for *E*, at least in the circumstances, and given determinism (Mackie [1965]).

To be viable, this idea needs careful expression. Start with the observation that 'the circumstances' must specify the *other* causes with which *C* combines

to bring about *E*. That leads to the suggestion that what is key is that the set of causes of an event *E* should *collectively suffice* for that event, but should do so *non-redundantly*: no proper subset should suffice for *E*. Then this set had better not include *all* the causes of *E* occurring at *any* time, since later ones will render earlier ones redundant, and vice versa. So amend, requiring that the set of causes of *E* that *occur at some given time* (before *E* occurs) non-redundantly suffice for *E*. That amendment still does justice to the guiding idea. What remains is to say what 'suffice' means.

A first pass: a set S of events suffices for (later) event *E* just in case it would violate the fundamental laws for the events in S to occur but for *E* to fail to do so. That won't do, since in general it will be possible for the events in S to occur jointly with some other 'inhibiting' events that act so as to *prevent* the occurrence of *E*. In Figure 3, it obviously doesn't follow from the fact that **C** fires, together with the 'neuron laws', that **E** will fire—for **A** could have fired, and if so would have prevented **E** from firing. (Some will say that we should simply count **A**'s *failure to fire* as itself a cause of *E*. For reasons why this is a bad idea, see Hall and Paul [2002].)

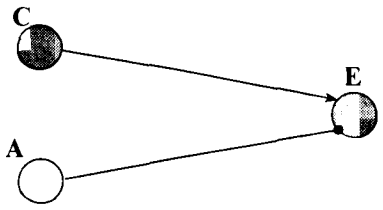

Figure 3

A better idea is to say that S suffices for *E* just in case, were the events in S to occur *without any interference*, *E* would occur. If we agree that such interference would require the occurrence of some other, contemporaneous, event, then we can simplify: a set S of events occurring at some time *t* suffices for (later) event *E* iff, were the events in S the *only* events occurring at *t*, *E* would (still) occur. (A counterfactual construction, yes—but do not be fooled into thinking that this approach must therefore be similar in spirit to the counterfactual analyses we will consider in the next section.) Calling a set *minimally* sufficient just in case it is sufficient, but no proper subset is, we thus arrive at the following *updated regularity account*: *C* is a cause of *E* iff *C* belongs to a set of contemporaneous events that is minimally sufficient for *E*.

The account enjoys an immediate success, for it neatly circumvents a problem long thought to be fatal to any regularity account. Consider Figure 4. **C** fires, sending signals to **B** and **D**, both of which fire; **E** fires upon receiving the signal from **D**. *B* does not cause *E*. But surely *B* is *sufficient* for *E*, at least in the circumstances: from a suitable specification of these circumstances,

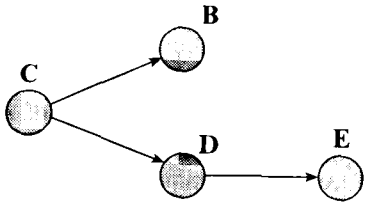

Figure 4

together with the relevant laws, it follows that, if **B** fires at t, then **C** fires a little earlier, and so **D** fires at t, and so **E** fires a little later. Therefore, regularity accounts must inevitably count B a cause of E, mustn't they?

Not ours. Let t be the time at which **B** and **D** fire, and consider the two relevant sets, $\{D\}$ and $\{B\}$. The first is, clearly, minimally sufficient for E: for if D alone had occurred at t, then E still would have occurred. The second, equally clearly, is not: if B alone had occurred at t, then E would *not* have occurred. Where is the problem? Do not say: 'The problem is that the counterfactual situations being considered violate the laws: for how (e.g.) could **B** fire at t without **D** also firing then?' The question is foolish twice over. First, in the dialectical context—one in which regularity accounts are being considered as rivals to counterfactual accounts—the question applies equally well to the opposition. (How could **B** fail to fire at t, without **D** also failing to fire?) Second, it applies rather poorly, since it is patently consistent with the dynamical laws at work here that at t, the four neurons are connected as shown, and only **B** is firing. If you are having trouble seeing this, just consider that the world could *start out*, at t, in just such a state (thanks here to Tim Maudlin). And if it did, it would be perfectly obvious and determinate how it would evolve, in accordance with the dynamical 'neuron' laws.

This success notwithstanding, we will see in the sections ahead that our updated regularity account will need amending if it is to have any hope of success. But in the simple form displayed here, it is an excellent and useful example of the type.

On to the examples.

5 Early pre-emption

There are several different varieties of redundant causation where, intuitively, an actual cause C of some event E is accompanied by *backups*, poised to bring about E in C's absence. The best known sort of case involves *early pre-emption*, in which a potential causal chain is interrupted by the causal chain that brings about the effect. Figure 1 provides a canonical example. Neurons **C** and **A** fire simultaneously: each neuron initiates a causal process that, were it to go to

completion, would stimulate neuron **E**. However, *C* initiates a side process that interrupts the process initiated by *A*. Intuitively, *C* is a cause of *E* and *A* is not. *C* is a *pre-empting* cause, *A* a *pre-empted* backup.

Although it is perfectly clear that *C* is a cause of *E* and *A* is not, simple versions of counterfactual and regularity accounts say otherwise. A counterfactual analysis that identifies causation with dependence fails to count *C* a cause of *E*: for, if *C* had not occurred, *E* still would have occurred, caused by *A*. Simple regularity analyses, where *C* is a cause of *E* iff *C* (together with the laws) is sufficient (in the circumstances) for *E* are also inadequate: rightly counting *C* a cause, they also wrongly count *A* a cause. Our updated regularity account fares no better, since, focusing on the time at which **A** and **C** both fire, we find that $\{C\}$ and $\{A\}$ are both minimally sufficient for $\{E\}$. What to do?

Let's start by surveying some things *not* to do. One seemingly obvious solution is a non-starter in the context of an analysis of causation: individuate events in part by their causal origins. If the firing of **E** brought about by the pre-empting cause is *numerically different* from the firing of **E** that would have been caused by the pre-empted backup, because different in its causal origins, then *E*—the firing of **E** that *actually* occurs—does in fact depend on the pre-empting cause, and the pre-empted backup does in fact fail to suffice for it (van Inwagen [1978]). But such a strategy for addressing early pre-emption (and problems of pre-emption in general) is off the table. If we are trying to construct an analysis of causation, we cannot assume that we can distinguish 'ahead of time', as it were, the pre-empting cause from the pre-empted backup.

A second bad idea, more difficult to discern as such, begins with the seeming insight that part of what makes the example tick is that the process initiated by *A* fails to go to completion. A number of approaches in the literature attempt to exploit this fact (see e.g. Ramachandran [1997, [1998], and Ganeri *et al.* [1996], [1998]). Ramachandran gives a nicely compact statement of the common idea behind them:

> It seems true in all cases of causal pre-emption . . . that the pre-empted processes do not run their full course, as we might put it. For any pre-empted cause, *x*, of an event, *y*, there will be at least one possible event . . . which fails to occur in the actual circumstances *but which would have to occur in order for x to be a genuine cause of y*. (Ramachandran [1997], p. 273; italics in the original)

In our view, this is a mistake. Consider a slight variant on our example of early pre-emption (Figure 5). This diagram needs more careful description. First, let us stipulate that neurons can fire with different intensities, emitting signals with corresponding intensities. Second, let us stipulate that neuron **E** will fire iff it receives stimulatory signals of combined intensity 10 or greater. Third, **A** fires with intensity 10, and **C** with intensity 5. Finally, the inhibitory connection between **C** and **B** does not prevent **B** from firing, but just reduces

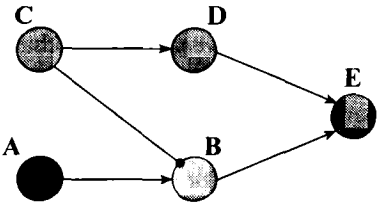

Figure 5

the intensity of its firing from 10 to 5. Since, as a result of the stimulatory signal from **C**, **D** likewise fires with intensity 5, **E** fires.

The case differs in certain respects from that depicted in Figure 1. In particular, *A* here is a cause of *E*; in fact, *C* and *A* are *joint* causes of *E*. But the similarities are more important. As in Figure 1, *E* fails to depend on *C*, one of its causes. But this failure of dependence does not obtain because some backup process, sufficient to bring about *E* all by itself, is *cut off*: rather, it obtains because some backup process, sufficient to bring about *E* all by itself, is *modified*, so that it is no longer sufficient in this way. Observe that this modification *does not prevent the process from going to completion.* Consequently, any attempt to handle the kind of pre-emption exhibited in Figure 1 by exploiting the fact that the *A*-process is cut short will fall apart when confronted with the minor variation exhibited by Figure 5. The counterfactual analyst is better advised to pursue some entirely different strategy for handling early pre-emption. We will consider two.

The first comes from Lewis [1973], and is built into the first of our three candidate counterfactual analyses: identify causation not with counterfactual dependence, but with the ancestral of counterfactual dependence. We then observe that in Figure 1 *E* depends on *D*, and *D* on *C*; likewise in Figure 5. This approach makes transitivity a non-negotiable feature of the account—a fact that some philosophers hold carries too great a cost (see McDermott [1995]; Hitchcock [2001]; for responses, see Lewis [2003], Hall [2000]). Note that, with minor tweaking of Figure 1 (left as an exercise), we can guarantee that *E* does not even counterfactually covary with *C*, so that a covariation account must also appeal to transitivity to handle this kind of pre-emption. (See Lewis [2003], which grants this point.)

The second approach is to try to show that *E* de facto depends on *C*. We already saw in section 3 how to do this for Figure 1. For Figure 5, focus on the path from *C* to *E* that goes through *D*. Hold fixed the off-path fact that **B** fires only with intensity 5; then, if **C** had not fired, **E** would have received only one intensity-5 signal and so would not have fired. Hence *E* de facto depends on *C*; hence *C* is a cause of *E*. A nice result—though, as we will see in the next section, it would be premature to declare victory.

Turn now to the challenges early pre-emption poses for a regularity account. If we focused exclusively on Figure 1, we might think that the problem is this: there are, at the time at which **A** and **C** fire, too many sets that are minimally sufficient for E, and in particular there are sets containing *non*-causes that are nevertheless minimally sufficient for E (namely, the set $\{A\}$). But Figure 5 shows that there is more to the story. For in that example there is a unique set minimally sufficient for E—the set $\{A\}$. The problem is that C, one of the causes of E, does not belong to it. So it would be more accurate to characterize the problem common to both cases in the following way. There is a cause of E that fails to belong to a unique minimally sufficient set for E (either because there is no *unique* set, or because there is, but the cause does not belong to it). On the other hand, at the time at which **B** and **D** fire, there is no such problem: $\{B, D\}$ is the unique minimally sufficient set for E, and it contains all and only the causes of E occurring at that time. That suggests an approach on analogy with Lewis's original strategy: pick out a condition that is meant to be only a *sufficient* condition for causation, and then extend it by 'taking the ancestral'. In the case of the counterfactual analysis, the condition sufficient for C to cause E is for E to depend on C. In the case of our regularity analysis, we can take the sufficient condition to be the following: C is a cause of E if, at the time of C's occurrence, there is a unique set of events minimally sufficient for E, and C belongs to it. We could then handle the examples in Figures 1 and 5 by taking causation to be the *ancestral* of this relation. For in both figures D belongs to a unique set minimally sufficient for E, and C belongs to a unique set minimally sufficient for D. Again, this approach weds our regularity account to transitivity.

In summary, it appears that a regularity account needs to handle early pre-emption by what is essentially an appeal to the transitivity of causation. It appears that counterfactual accounts have more options: they can appeal to transitivity, or they can appeal to the more complicated kind of 'dependence' featured in de facto dependence accounts. Let us now raise the ante by introducing a much more stubborn class of counterexamples.

6 Late pre-emption

Suzy and Billy, two friends, both throw rocks at a bottle. Suzy is quicker, and consequently it is her rock, and not Billy's, that breaks the bottle. But Billy, though not as fast, is just as accurate: had Suzy not thrown, or had her rock somehow been interrupted mid-flight, Billy's rock would have broken the bottle moments later. This case, like those reviewed in the last section, features a cause of an event that is accompanied by a backup, sufficient to bring about the effect in the cause's absence. But, unlike those cases, the actual cause fails either to interrupt the backup process, or to interact with it in such a way as to

render it insufficient for the effect. Consequently—and this is the crucial feature that distinguishes these cases of so-called 'late pre-emption' from the cases of early pre-emption just considered—*at no point in the sequence of events leading from cause to effect does there fail to exist a backup process sufficient to bring about that effect.*

Consider Figure 6. The signal from **C** reaches **E** just before the signal from **A**; we represent this fact by drawing the arrow from **A** so that it does not quite extend to **E**. Pick any point before **E** fires. Focus on the event occurring at that time which is a part of the passage of the signal from **C** to **E**. If that event had not occurred, **E** would have fired all the same. What's more, that event is not part of a unique set minimally sufficient for *E*. (It is, rather, a member of one of *two* sets each minimally sufficient for *E*.) So the strategy of taking the ancestral won't help the counterfactual or the regularity analysis. But before considering what other strategies might be of use, we need to get clear on what is and what is not essential to such cases.

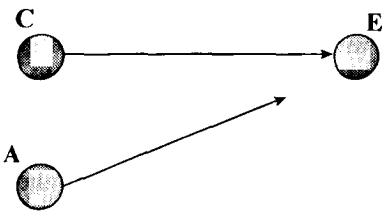

Figure 6

To begin, the name 'late pre-emption' is unfortunate. It hearkens back to Lewis [1986*b*], whose canonical example featured an effect itself that acted so as to interrupt the backup causal process, as in Figure 7. (Imagine here that **A** and **C** fire simultaneously.) Here, the inhibitory signal from **E** prevents **D** from firing; if it had not done so, the signal from **A** would have caused **E** to fire. Take this example as paradigmatic, and you will think the term 'pre-emption' is appropriate, since there is an obvious sense in which the backup process is *cut short*. But in Figure 6 we can stipulate that, after **E** fires as a result of the

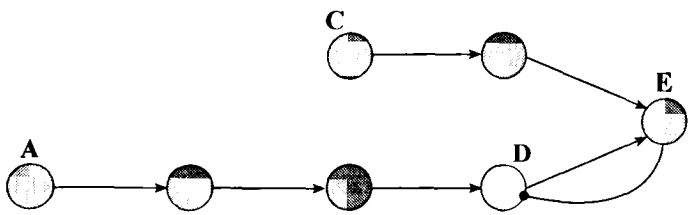

Figure 7

signal from C, it fires a *second* time as a result of the signal from A. There is thus no sense whatsoever in which the process initiated by *A* fails to go to completion. Still, it is too late to go legislating a linguistic change; hence we will stick with the term 'late pre-emption'.

The focus on examples like Figure 7 has done more than terminological damage. For notice that, in the counterfactual circumstance in which C does not fire, there is an event in the *A*-process—namely *D*, the firing of D—that does not *actually* occur. That can make approaches that exploit the presence of such 'gaps' seem attractive (see Ganeri *et al.* [1996]; Ramachandran [1997]). We already saw in the last section that their attractiveness dims once one is clear about what is essential to early pre-emption. So too here: cases of late pre-emption like Figure 6 show that the pre-empted causal process need not be interrupted by the prevention of an event. Consequently, there need be no absent event available for analyses to target as evidence of such an interruption.

An extremely natural thought, at this point, is to fix on the observation that without Suzy's throw the bottle would not have shattered *just when it did*. Likewise, in Figure 6, without *C*, E would not have fired as early as it did. Perhaps these facts are somehow crucial to distinguishing Suzy's throw as the sole cause of the shattering, and *C* as the sole cause of the first firing of E.

One could try to exploit this observation by building an extreme sensitivity to the time and perhaps manner of occurrence into the individuation conditions for events (or, making use of what is a well-entrenched bit of technical terminology, to treat events as 'modally fragile'; see Lewis [1986*b*]). Thus, the shattering of the bottle that actually occurs—that very event—would not have occurred without Suzy's throw; rather, a qualitatively very similar but *numerically different* shattering would have occurred. Or one could build the sensitivity to time into the kind of counterfactual dependence that one takes to suffice for causation. For example, we could take *C* to be a cause of *E* if, had *C* not occurred, *E* would not have occurred exactly when it did—leaving it open whether the consequent holds because *E* does not occur at all, or because it occurs at a different time (Paul [1998*b*]). Lewis's [2003] 'influence' account is yet another attempt to exploit the fact that, in cases of late pre-emption, the effect is differentially counterfactually sensitive to the cause, as opposed to the pre-empted backup.

The fatal problem, across the board, is that there is no such fact. (For other problems, see Lewis [1986*b*] and Schaffer [2001].) It is perfectly easy to construct late pre-emption examples in which, had the cause not occurred—or, indeed, had any of the events connecting the cause to the effect not occurred—the effect would have occurred at exactly the same time, and in exactly the same manner. In Figure 6, for example, suppose that the signal from C exerts a slight retarding force on the signal from A. Pick any point before this signal

from C reaches E, and ask what would have happened if, at that time, the signal had been absent. Answer: the signal from A would have accelerated, and we can stipulate that it would have accelerated enough to reach E at exactly the time at which the signal from C in fact reaches E.

But might not *E* at least counterfactually covary with *C*, or with one of the intermediate events? There's no reason why it should. For example, let us stipulate that the stimulatory signals from C and A are different. The one from C is a 'special' signal, the one from A an 'ordinary' signal. Neuron E can be stimulated to fire by a special signal only if that signal's physical characteristics are just right, and only if the signal reaches it at just the right time. (E is not so finicky when it comes to ordinary signals.) What's more, the power the signal from C has to delay the signal from A is exquisitely sensitive both to the physical characteristics of the C-signal and to the time lag between the two signals. Piled on top of one another, these stipulations give the desired result: for if anything about *C* had been different—for that matter, if anything about *any* of the events leading from *C* to *E* had been different—then, in the first place, the C-signal would no longer have been capable of stimulating E, and, in the second place, the A-signal would no longer have been delayed, and so would have stimulated E to fire at exactly the time and in exactly the manner it does.

We anticipate the usual complaints—'The case is too arcane'; 'Our intuitions are not reliable when confronted with a case that violates so many of our ordinary presuppositions about how the world works'; and so on. Such objections are silly. Of course the case is convoluted, but the convolutions concern respects that are *irrelevant* to an evaluation of its causal structure. It will pay to see why. Accordingly, consider Figure 8. What we have depicted here is a duplicate of part of what is depicted in Figure 6; we have simply omitted A, and the signal traveling from it towards E. Otherwise, matters are *exactly the same*: C fires in just the same way; the signal traveling from it to E has exactly the same physical characteristics and reaches E at exactly the same time; E is finicky in just the same respects. Now forget about Figure 6, and point your intuitions firmly in the direction of this pared-down example. Is there any doubt whatsoever that *C* causes *E*? Of course not. So return to Figure 6, taking that firm intuitive judgement with you, and observe that we arrive at Figure 6 merely by adding what is clearly a *causally idle* backup process to Figure 8. How could this addition possibly make a difference to the fact that *C* causes *E*? And, even supposing it did, should we then reason that *E* has no causes at all, or that it is somehow caused by *A*? Of course not. We conclude that there

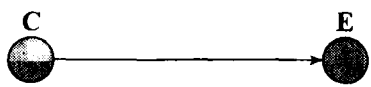

Figure 8

is no serious challenge to the intuitive verdict about Figure 6 that C is a cause of E. But, since E does not counterfactually covary with C at all, the 'sensitivity' strategies we are considering for handling late pre-emption fail here.

However, the grounds we gave for supporting our intuitive verdict about Figure 6 point to an alternative, and quite different, strategy. Many philosophers have been attracted to the idea that the causal structure of a process is *intrinsic* to it (Lewis [1986*b*]; Menzies [1996]; Hall [2002]). Arguably, some such principle is at work in guiding our intuitive judgements about Figure 6, and about cases of late pre-emption quite generally. (Notice how it was appealed to in the argument of the last paragraph.) It is difficult to state this thesis in a way that is both precise and defensible. We will simply short-circuit these issues by adopting the formulation defended by Hall [2002]. (Hall shows there why simpler versions won't work.) The key idea is that, if we pick out a comprehensive structure of events—namely, a structure consisting of an event together with all of its causes back to some earlier time—then the causal characteristics of that structure will be fixed solely by its intrinsic character, together with the governing laws.

6.1 Intrinsicness

Let S' be a structure of events consisting of event E', together with all of its causes back to some earlier time t. Let S be a structure of events that intrinsically matches S' in relevant respects, and that exists in a world with the same laws. Let E be the event in S that corresponds to E' in S'. Let C' be some event in S' distinct from E', and let C be the event in S that corresponds to C'. Then C is a cause of E.

Some terminology will help here. Call a structure of events an *E-blueprint* just in case it consists of some event E' that intrinsically matches E, together with all of the causes of E' back to some earlier time. The intrinsicness thesis says that, if C and E belong to a structure of events that intrinsically matches some nomologically possible E-blueprint, then C is a cause of E.

If this thesis is right, it goes a long way towards explaining how our intuitive judgements about cases of late pre-emption work. For, plausibly, we can arrive at any such case by beginning with a case of perfectly ordinary, garden variety causation, and then adding details to it that concern matters extrinsic to the processes that bring about the target effect. Thus, we can arrive at Figure 6 by starting with Figure 8 and adding such details; we can arrive at the case of Billy, Suzy, and the bottle by starting with a situation in which Suzy alone throws a rock at the bottle and then adding extrinsic details (Billy and his throw); and so on. If, as our intrinsicness thesis states, such extrinsic changes make no difference to the causal structure of the process we begin with, then it is no surprise that cases of late pre-emption evoke such clear and

firm intuitive judgements: we 'see' in them, as it were, perfectly ordinary cases of causation.

How to turn this insight (if such it be) into a strategy that reductive analyses can pursue for handling late pre-emption is another question. Hall [2002] gives a detailed treatment of this issue, of which we will reproduce the main points. The key idea is to begin with an analysis that gets easy cases right—and does not get cases of late pre-emption *wrong*—and to extend it by appeal to the intrinsicness thesis. As an illustration, suppose we have come up with a provisional analysis that gets Figure 8 right: it successfully identifies *C*, together with the events constituting the passage of the stimulatory signal to **E**, as being all the causes of *E*. But it falls silent about Figure 6. Still, Figure 6 contains within it a perfect intrinsic duplicate of the events depicted in Figure 8. According to our intrinsicness thesis, then, *C* in Figure 6 is a cause of *E*. More to the point: the combination of our provisional analysis and our intrinsicness thesis *entails* that *C* is a cause of *E*.

What emerges is a general strategy for handling cases of late pre-emption. Confronted with such a case, the first step is to 'extract' from it a simpler case, by considering what would have happened if no backup processes had been present. This simpler case will serve as blueprint for the effect in question. If we can produce an analysis that adequately describes every such blueprint, then augmenting it with our intrinsicness thesis will enable it to handle any case of late pre-emption, as well.

There are a number of challenges to developing this 'blueprint strategy' that we must simply note and pass over (but see Hall [2002] and Hall and Paul [2002] for details). The qualification 'in relevant respects' that appears in the statement of the intrinsicness thesis is essential (and obviously needs to be explained); for pre-empted non-causes such as Billy's throw can still belong to event structures that are similar in *some* respects to appropriate blueprints. It turns out to be tricky to develop an analysis that will not only identify all the causes of *E* in a simple case where no backups are present, but will also identify them as *being* all the causes; but there is no way to implement the blueprint strategy without satisfying this demand. And the blueprint strategy will not, unfortunately, cover all the kinds of *early* pre-emption we saw in the last section: for example, it falls silent about Figure 5. (The reason is that, if we try to extract a 'blueprint' by removing the inhibitory connection between **C** and **B**, we end up with an event structure too intrinsically dissimilar from the relevant structure of events in Figure 5.) So the blueprint strategy must, apparently, be used in conjunction with an appeal to transitivity in order that all the cases of pre-emption so far discussed can be accommodated.

A final problem merits more detail. The blueprint strategy turns out to *conflict* with the attractive thesis that counterfactual dependence is sufficient for causation. For there are situations in which *E* depends on *C* only because *C*

prevents something from happening which, had it happened, would in turn have prevented *E*—as in Figure 9. Here, if **C** had not fired, then the signal from **B** would have stimulated **D** to fire, which in turn would have prevented **E** from firing. This kind of 'dependence by double prevention' strikes many as a kind of causation, but if so it is a kind of causation that cannot survive changes extrinsic to the events that exhibit it. For compare Figure 10, which we arrive at simply by excising **B**, and its connection to **D**. Here, the non-existent neuron **B** quite obviously poses no threat to the firing of **E**; *C* cannot therefore count as a cause of *E* in virtue of counteracting this threat. But if we pick out, in Figure 9, a structure consisting of *E* together with all of its causes, then, *even if these are taken to include C*, we will find that structure duplicated perfectly in Figure 10. So if *C* in Figure 9 is a cause of *E*, then either our intrinsicness thesis is false or, implausibly, *C* in Figure 10 is likewise a cause of *E*. (But see Hall [2003] for a third option: perhaps our intrinsicness thesis is true of one central *kind* of causation.)

This is an important result: it shows that the core thesis of counterfactual analyses conflicts with the intrinsicness thesis, and so the blueprint strategy for

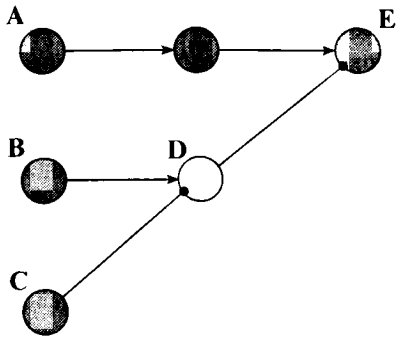

Figure 9

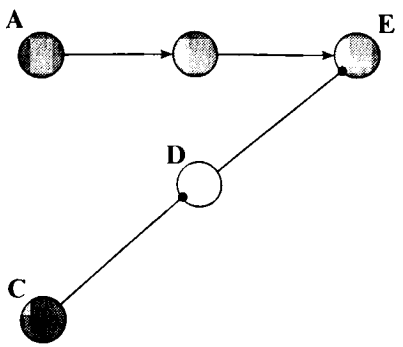

Figure 10

handling cases of late pre-emption that we have just articulated is not available to such analyses. By contrast, Figure 9 reveals no conflict between the intrinsicness thesis and our candidate regularity analysis; for, since $\{A\}$ is the unique set minimally sufficient for E, C does not even qualify as a cause of E, according to that analysis. (See Hall [2002] and [2003] for more discussion, and the latter paper for a more sophisticated regularity analysis that appears capable of taking full advantage of the blueprint strategy.)

We do not mean this to be some serious criticism of counterfactual analyses—one could just as easily blame the intrinsicness thesis for the conflict. It's rather that what we have here is a dramatic fork in the road, a point at which counterfactual and regularity analyses must clearly diverge in the approach that they take to handling pre-emption, and diverge in a way that will have far-reaching ramifications for the picture of causation they ultimately settle on.

It may seem to readers that the cost of the blueprint strategy is too high. (And the price will rise at the end of this section, where we examine an apparent counterexample to our intrinsicness thesis.) But don't hide your philosophical checkbook just yet: for no alternative stands out as clearly superior to it. What alternatives does one find in the literature? Well, one finds a package of alternatives that try to exploit the fact that, in cases of pre-emption, the backup process fails to go to completion. But there is no such fact to exploit. Again, one finds a package of alternatives that try to exploit the fact that the effect is differentially counterfactually sensitive to its cause. But this fact does not exist, either. What is left?

We know of one more approach that needs serious consideration, which is to try to extend the de facto dependence treatment of early pre-emption discussed above to cover cases of late pre-emption as well. For example,

> Even without considering fine differences in the way Suzy throws her rock, or in the way the bottle shatters, it can be shown that there is an active route from Suzy's throw to the bottle's shattering. This is revealed by the following . . . counterfactual: *given* that Billy's rock did not hit the bottle, if Suzy had not thrown, the bottle would have remained intact throughout the incident. (Hitchcock [2001], p. 289)

This approach promises a uniform treatment of both early and late pre-emption, unlike the bifurcated approach just considered, which sometimes needs to appeal to the intrinsicness thesis, and sometimes to transitivity. It can accommodate causation by double prevention. It is not wedded to transitivity, and so may avoid the apparent counterexamples to that principle (but see below for a qualification). And the key idea behind the approach has a great deal of intuitive appeal. Granted, in Figures 1, 5, and 6 E does not *need* C in order to come about—for in C's absence, something else would have sufficed. But there still seems a clear intuitive sense in which, *as things in fact play out*, it is at least in part C that meets the needs of E. We can think of the concept of

de facto dependence as a device for making this intuitive idea precise. (This way of laying out the motivation for the account we borrow from Yablo [2003].)

However, as you can by now perfectly well predict, there is trouble brewing. We will consider four problems.

First, it is far too obscure what truth conditions govern the kind of counterfactual that Hitchcock would have us consider. Hitchcock suggests that we 'hold fixed' the fact that Billy's rock never touches the bottle. To what value of what variable does this fact correspond? None, as it seems. Perhaps the structural equations variant of a de facto dependence approach is bumping up against its limits here, as we apparently need some *other* story about how it is that this fact gets selected. (For one version of such a story, see Yablo [2003]. For an attempt to deal with this issue from within the structural equations approach, see Pearl [2000], Ch. 10.)

Waiving this issue, is Hitchcock's conditional *true*? Remember that we are not to evaluate this conditional by, for example, supposing that Billy likewise fails to throw his rock—for that is not to find actual factors and hold them fixed, but rather to find actual factors and alter them. (Proceeding in this way would also qualify *Billy's* throw as a cause.) No, the counterfactual situation envisaged is one in which Suzy does not throw but Billy does throw, and with just as deadly an aim. It's completely unclear how such a situation is supposed to play out, *given* that Billy's rock must somehow fail to strike the bottle. Note that this problem is not completely separate from the one just discussed: for if what is to be held fixed are the values of some variables in some appropriately constructed causal model, then the details of that model tell us exactly how to construct the counterfactual scenario. We plug in the values that are to be held fixed; we plug in the counterfactual value for the *C*-variable; and then we calculate, using the appropriate functional dependence relations, the values for all other variables, in particular the *E*-variable. But this method is not available, in the case at hand. So as it stands, we are left adrift.

Worse than adrift, really, since perfectly plausible ways of constructing the needed counterfactual scenario have the result that the bottle *does* shatter. Suppose the bottle is perched on a post. Then one way it could come about that Suzy does not throw, that Billy does throw, and that Billy's rock somehow fails to strike the bottle, is for a gust of wind to knock the ball off the post before Billy's rock reaches it—in which case, fragile thing that it is, it shatters upon hitting the ground.

Third, Hitchcock's approach, if it works at all, does so merely because of inessential features of the case. Recall that Hitchcock tells us to hold fixed the fact that Billy's rock does not hit the bottle. But who says it doesn't? Imagine that, in the brief interval of time between the shattering of the bottle and the arrival of Billy's rock, a genie quickly reconstructs the bottle, so that Billy's

rock shatters it a second time. That change does not in any way undermine the original verdict that Suzy's throw alone is a cause of the (first) shattering. In fact, we built this variant into our Figure 6, for we stipulated there that the signal from **A** causes **E** to fire a second time, moments after the signal from **C** reaches it. So what exactly is to be held fixed, here?

Finally, even if we waive all of these problems, the account makes it too easy for *C* to be a cause of *E*. (Note that what follows spells trouble equally for the de facto dependence account of early pre-emption.) Consider a situation in which *C* initiates two processes, one of which threatens to prevent *E*, and the other of which counteracts that threat, as in Figure 11. *E* occurs. Obviously, *E* does not depend simpliciter on *C*. Nor ought *C* to count as a cause of *E* simply because it tries to prevent *E* (by way of *B*), but manages to sabotage its own efforts (by way of *D*). (For a bizarrely contrary view, see Lewis [2003]; for an amusing rebuttal, see Yablo [2003].) But if we hold fixed the firing of **B**, then, if **C** had not fired, *D* would not have prevented *B* from causing **F** to fire, and so **E** would not have fired. This problem generalizes well beyond the boundaries of Figure 11. Take any case where *C* threatens to prevent *E* via one process, but simultaneously initiates a second process that counteracts this threat: if we hold fixed the presence of the threat initiated by *C*, then if *C* had not happened that threat would not have been undone, and so *E* would not have happened. So, in general, *E* de facto depends on *C* in such a case.

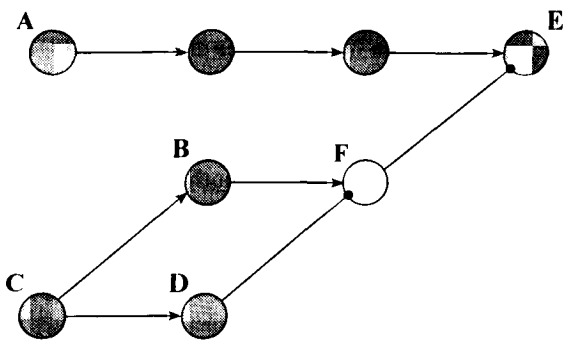

Figure 11

(Such examples are not simply an embarrassment for the de facto dependence account. They also spell trouble for transitivity, at least when combined with the claim that dependence suffices for causation: observe there is a two-step chain of dependence linking *C* to *E* in Figure 11. That means, incidentally, that it's a bit of false advertising for de facto dependence accounts to claim that it is an advantage of their approach that they are not wedded to transitivity. See Hall and Paul [2002] for more discussion.)

How do de facto dependence accounts respond to this problem? Yablo [2003] takes it quite seriously, and attempts to construct principles for determining what can legitimately be held fixed that will, for example, forbid us from holding *B* fixed, in Figure 11. The key idea is that *C* is disqualified from being a cause of *E* if it merely introduces new needs that must be met, in order for *E* to occur; *C* in Figure 11 does so by making it the case that **D** must fire, in order for **E** to fire. Unfortunately, it's also true of Figure 5 that *C* merely introduces new needs, since it likewise makes it the case that **D** must fire, in order for **E** to fire. Yablo's maneuver thus utterly fails to show why *C* in Figure 5 is a cause of *E* whereas *C* in Figure 11 is not.

Hitchcock [2001] is somewhat more dismissive, appealing to what are essentially pragmatic factors in order to explain away the strong intuition that *C* does not cause *E*. Considering a case with the same structure as Figure 11, he claims that the relevant possibility—in our case, the possibility in which **C** does not fire, but **B** nevertheless *does* fire—is 'just too far-fetched' to be taken seriously. That strikes us as unhelpfully vague, but at any rate there is a more obvious problem: Hitchcock's reasoning applies equally well to Figures 1, 5, and 6. Take the first of these: why is it not just as 'far-fetched' a possibility that **C** fires, but that **B** nevertheless fails to fire? No reason, as far as we can see. But then Hitchcock ought to conclude that it should, at the very least, 'sound odd' to say of *C* in Figure 1 that it is a cause of *E*. On the contrary, of course, it sounds exactly right.

Taken together, these four problems seem to us quite serious, rendering the prospects for a successful de facto dependence account of causation uncertain at best.

Let us now try to make matters worse—for both the de facto dependence and blueprint strategies. Can we think of some further refinement to our late pre-emption counterexample that renders both approaches ineffective—our most virulent example of late pre-emption yet? Perhaps. We will present a candidate for such an example, although we will see that the price for its greater intractability is a certain loss of intuitive clarity, and that the two approaches may yet have some room for maneuver.

Suppose that neurons are characterized by two distinct parameters: call them 'color' and 'intensity'. In Figure 12, neuron **C** can fire with various

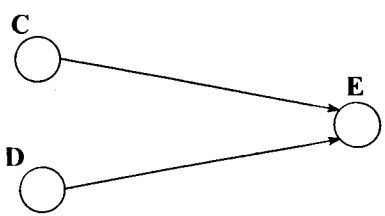

Figure 12

different colors and in various different intensities, as can **D**. Here are the laws that govern this particular arrangement: first, if **C** does not fire and **D** does, then, regardless of the color or intensity of the signal from **D**, **E** will fire when that signal reaches it. Second, if **D** does not fire but **C** does, then **E** will fire iff **C** fires in a particular shade of green. Third, if **C** fires with a particular intensity and in some other shade than this 'triggering' shade of green, then **E** will not fire, regardless of whether and how **D** fires. Fourth, if **C** fires in some intensity other than this 'inhibiting' intensity, then **E** will fire iff either **C** fires in the triggering shade of green, or **D** fires. Fifth, the way that **E** fires is never sensitive to the character of the stimulatory signal that causes it to fire. We have thus covered all of the cases except one: **D** fires, and **C** fires in the triggering shade of green (represented by cross-hatching), and with the inhibiting intensity. Let us suppose that in this case, **E** fires, as shown in Figure 13.

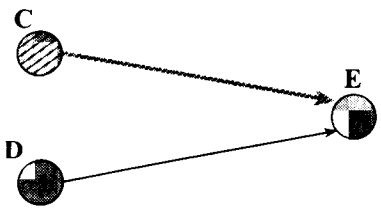

Figure 13

Before considering the causal structure of this last case, let us draw out the most natural causal gloss to put on the various *other* behaviors that this system can exhibit. Clearly, if **D** alone fires, then the right thing to say is that *D* causes *E*. Likewise, if **C** alone fires in the triggering shade, then the right thing to say is that *C* causes *E*. If **D** fires, and **C** fires with the inhibiting intensity (but not in the triggering shade), then **E** will not fire, and the right thing to say is both that *C* prevents **E** from firing, *and*—what is subtly different—that *C* prevents the *D*-signal from *causing* **E** to fire. If both **D** and **C** fire, and **C** fires neither in the triggering shade nor with the inhibiting intensity, then the right thing to say is that *D* alone causes *E*. And finally, if **C** fires in the triggering shade but not with the inhibiting intensity, and **D** also fires, then it seems that the best thing to say is that *E* is symmetrically overdetermined by *C* and *D*. The upshot is that the labels 'triggering shade' and 'inhibiting intensity' seem entirely appropriate; that is, it seems that we can assert as causal generalizations that a signal from **C** with the inhibiting intensity acts to prevent any *D*-signal from causing **E** to fire, and that a signal from **C** with the triggering shade is capable of itself of causing **E** to fire. What would seem most natural is to hold that these generalizations extend even as far as Figure 13, so that (i) the *C*-signal prevents the *D*-signal from causing **E** to fire (because the *C*-signal has

the inhibiting intensity); and (ii) the *C*-signal itself causes **E** to fire (because it has the triggering shade).

Let us take this as the correct gloss, while recognizing that there is certainly more room for dispute here than there was in our earlier cases of early and late pre-emption. Then we have a particularly nasty example of late pre-emption. Pick any event in the *C*–*E* chain, and ask what would have happened if it had not occurred: the answer is that **E** would have fired all the same. Similarly, focus on any time before the time at which **E** fires, and you will find at least two distinct minimally sufficient sets for *E*, one containing the appropriate event in the *C*–*E* process, the other containing the corresponding event in the *D*–*E* process. Try to fall back on the intrinsicness thesis, and you will certainly get the result that *C* is a cause of *E*, but you will also get the result that *D* is a cause of *E*; for the structure of events connecting *D* to *E* perfectly matches a blueprint in which *C* and the signal it generates are absent. And an appeal to de facto dependence seems equally futile: hold fixed whatever you like in the *D*–*E* process; without *C*, **E** would have fired all the same.

We see two possible responses. First, a defender of a covariation account can be expected to pipe up at this point, observing that, while *E* does not depend on *C*, it *is* the case that, had **C** fired differently—specifically, with the same intensity but in a different color—then **E** would not have fired. So there is some limited 'influence' of *C* on *E* here. *D*, by contrast, appears to exert no such influence on *E*.

But of course that cannot be the whole story: we already know how to construct late pre-emption cases that a covariation account cannot handle at all. In fact, we can simply tinker with the case before us: add to the environment various 'color detectors' that check to see whether the *C*-signal is in the appropriate triggering shade—and if not, quickly act to guarantee that **E** fires by other means. Such backup processes can obviously erase any influence that *C* (equally: any of the events in the process it initiates) has on *E*. Still, at this point a defender of a de facto dependence approach might try to appropriate elements of the covariation account, claiming that, if we hold fixed not merely the facts about the *D*-process, but *also* the fact that no backup processes in the environment are ever engaged, then it will be the case that, if **C** had fired differently (i.e. with the same intensity but a different color), then **E** would not have fired. There thus might be some hope for wedding the de facto dependence approach to the covariation approach.

A second strategy is to aim for more subtlety in both the statement and the application of the intrinsicness thesis. Recall that we stated that thesis solely as a constraint on *causation*. But does not some such thesis also govern the way we think about *prevention*? Suppose we have a situation in which two processes intersect, in such a way that the first prevents the second from bringing about some effect; then if the original intrinsicness thesis is plausible, it should also

be plausible that an intrinsic duplicate of this structure should share its causal characteristics, and so should likewise be a structure in which one process prevents the other process from bringing about some corresponding effect. That some such thesis guides our thinking about prevention would help explain why we arrive at the causal judgement we do about Figure 13: we note that the intrinsic character of the two intersecting processes is relevantly similar to that displayed in Figure 14, where **C** fires with the inhibiting intensity but not in the triggering shade:

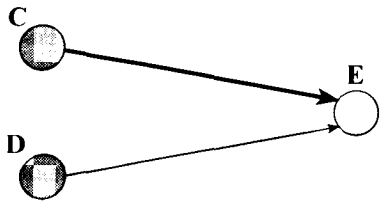

Figure 14

But even if this is right, we don't yet have a solution to our problem; rather, we just have an apparent conflict between an intrinsicness thesis that governs causation and an intrinsicness thesis that governs prevention. Apparently, there are contradictory ways to 'project' causal structures from simple situations into complex situations that feature intrinsic duplicates of those simple situations. For example, we could appeal to similarity in intrinsic character to project the causal structure exhibited in Figure 15 into Figure 13, and thus arrive at the judgement that *D* causes *E*. Or we could appeal to intrinsic similarity to project the causal structure of Figure 14 into Figure 13, thus arriving at the judgement that *C* prevents *D* from causing *E*. A strategy for using the idea that causal structure is intrinsic to the processes that exhibit it, that will succeed in handling not only ordinary cases of late pre-emption such as Figure 6, but also virulent cases of late pre-emption such as Figure 13, must, it seems, appeal to some overarching principles to resolve conflicts such as these. We will not pursue this matter further, but see Hall [2002] for some discussion.

Figure 15

7 The lessons of redundant causation

Let us try to summarize the main lessons that cases of causal pre-emption appear to teach. A good way to begin is by noticing a general strategy for constructing counterexamples that emerges from the foregoing discussion. Suppose you have proposed necessary and sufficient conditions for causation that successfully identify all the causes of some event E in some reasonably simple example. The strategy for constructing a counterexample to your position is to try to come up with a situation just like the situation your analysis gets right—except in respects extrinsic to the causal processes leading to E—but where these extrinsic changes either introduce non-causes that meet your conditions, or make it the case that some of the genuine causes do not. Focus on the latter option. Pick some candidate cause C of E, and try to tinker with the environment so as to make it the case that C no longer bears what you take to be the definitive relation to E. If one can eliminate this relation—and do so merely by means of suitably extrinsic modifications to the environment of the processes connecting C to E—then one will almost certainly have a counterexample to the necessity of your analysis, *since such extrinsic modifications will not undermine the intuitive verdict that C remains a cause of E.*

Perhaps the best way to effect such modifications is to introduce some alternative process aimed at the effect, and sufficient to bring it about in C's absence. That the alternative process has this capacity will likely guarantee either that (i) its constitutive events will also bear the definitive relation to E (this is what happens to a simple regularity account, where we end up with more than one set of events minimally sufficient for the effect), or (ii) C no longer bears the definitive relation to E (this is what happens to a simple counterfactual account, where the presence of the backup processes destroys the dependence of E on C). Do not be distracted by the fact that (i) yields a counterexample to the sufficiency of the analysis, whereas (ii) yields a counterexample to its necessity; for we can almost always convert the former kind of counterexample into the latter, by making the target analysis more restrictive in some obvious way. For example, if our analysis says that a cause of E is any member of a set of contemporaneous events minimally sufficient for E, then standard cases of pre-emption will show that our analysis fails to provide a sufficient condition for causation. We can easily 'fix' this problem by making the analysis more demanding in the obvious way: let it say that a cause of E is any member of a *unique* set of contemporaneous events minimally sufficient for E. Now the very same cases will undermine the *necessity* of our analysis. In fact, we think it is most illuminating to take it for granted that we have started with an analysis restrictive enough to be relatively immune from counterexamples to its sufficiency; the strategy for refuting it is thus to start with a case where it works, and bring what are *intuitively irrelevant changes* into the environment

in such a way as to erase the conditions that the analysis identifies with causation.

We have seen that some easy ways of making this kind of modification leave it open that the targeted analysis succeeds, at least for the *late* stages of the process or processes leading to the effect; this is exactly what happens when we introduce alternative causes of *E*, and also introduce mechanisms by which the genuine causes cut them off. Again, the counterexample may leave it open that the idle backup processes are distinguished by the presence of 'gaps'—events that did not happen, but would have to have happened if those backups were to have brought about *E*. Finally, it may be that very subtle relations—in particular, of counterfactual dependence—still distinguish *C* from the events in the backup processes. The most typical cases of late pre-emption are like this, featuring backup processes that would have brought about *E* a little bit later, or in a slightly different manner, than its actual causes. But with enough care we can close all such loopholes. It is therefore unfortunate, if understandable, that so much of the literature has been devoted to developing 'solutions' to the pre-emption problem that take advantage of one or another of them.

A far better strategy is to try to find some relation between *C* and *E* that is guaranteed to be *stable* under modifications of the environment of the processes connecting *C* to *E*; for that is the one strategy that zeroes in on the crucial feature of the counterexamples that makes them tick. It is telling that the two approaches that appear to have any promise of being able to handle pre-emption—the blueprint strategy and the de facto dependence approach—both fix on relations that are stable in this way. It is pretty much obvious how the first of these approaches does so; for, if *C* and *E* belong to a structure of events that intrinsically matches, in relevant respects, some *E*-blueprint, then of course this fact will remain stable under modifications of the environment that leave the intrinsic character of this event structure unchanged (or unchanged in relevant respects). It is perhaps less clear that de facto dependence accounts delineate a relationship that remains stable under perturbations of the environment of the processes that exhibit it, and that is in part because it remains insufficiently clear how we are to evaluate the counterfactuals that feature in such accounts (especially in the context of late pre-emption). But bracketing that problem, it seems that relations of de facto dependence can remain stable in at least a wide range of cases. That is, it seems that, in at least typical cases, if *E* de facto depends on *C*, then, making modifications to the environment of the processes connecting *C* to *E*, it will *remain* the case that *E* de facto depends on *C*.

As partial vindication of this point, consider an arbitrary neuron diagram in which **C** is connected to **E** via a series of stimulatory signals and intervening neurons, and in which *C* is in fact a cause of *E*. On a de facto dependence account, this causal status is likely to be exhibited by the fact that, if we hold

fixed the pattern of neuron-firings off the path connecting **C** to **E**, we will find that if **C** had not fired then no stimulatory signal would have traveled this path to **E**, and so **E** would not have fired. But if this counterfactual is true, then it will typically *remain* true if we reconfigure the relationships and patterns of firings among these off-path neurons.

But we must highlight two caveats. In the first place, this result will not extend to examples in which *C* is straightforwardly connected to *E*, but in which the facts that need to be held fixed cannot simply take the form of facts about firings of neurons. Late pre-emption of the kind exhibited in Figure 6 is a prime example. In the second place, there are other situations in which reconfiguring the off-path neurons can introduce or erase relations of de facto dependence. If we start with Figure 11, we see that *E* de facto depends on *C* (holding *B* fixed). But, since the relevant path here is the **C–D–F–E** path, we can remove this de facto dependence simply by removing neuron **B** and the stimulatory connections linking it to **C** and **F**, as in Figure 16. That suggests a deeper diagnosis of part of the trouble facing de facto dependence accounts: they have almost—but not quite—fixed on a relation that is stable under environmental perturbations in the right way.

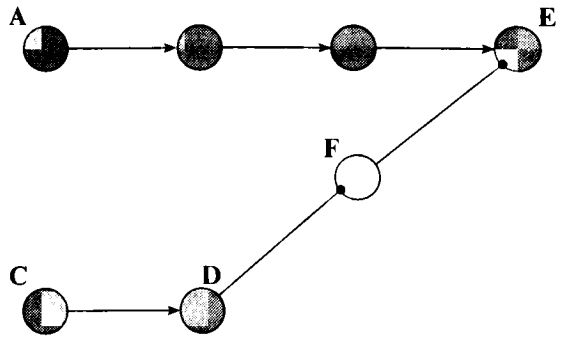

Figure 16

If we had to issue a summary judgement, we would say that, on balance, a regularity account that makes use of the intrinsicness thesis does a better job of handling causal pre-emption than does a de facto dependence account. Granted that there are important details of the regularity account that need to be worked out, and there are apparent counterexamples to the intrinsicness thesis (such as our Figure 13) that need to be dealt with in a principled way; still, the challenges facing a de facto dependence account—the only decent rival on the table—seem much more daunting. The foundations of these accounts remain deeply obscure, their ability to provide even reasonably precise truth conditions for the counterfactuals upon which they rely is highly

uncertain, and the serious counterexamples to them seem likely to be resistant to any principled treatment.

Still, it would be hasty to issue a firm judgement; for while we take ourselves to have covered in this essay the most important class of examples that drive debates about causation, there are other examples that a comprehensive treatment would need to explore—most especially, examples that highlight issues involving the causal status of omissions. The most significant of these feature relations of apparent causation by double prevention, in which an omission plays the role of an intermediate (as the failure of D to fire does in Figure 9, for example). This kind of relation is too widespread to be ignored, and there is a good case to be made that it would be a disaster to deny that it is any kind of causal relation (see e.g. Schaffer [2000*c*]). But, while it is clear that a counterfactual account can easily accommodate causation by double prevention, it is also clear that the sort of regularity account that is wedded to the intrinsicness thesis cannot (for further discussion, see Hall and Paul [2002]). So at the end of the day the balance sheet may look rather different: it may turn out that de facto dependence accounts, despite their struggles with pre-emption, provide the most attractive analysis of causation on offer. Then again, it may also turn out that the best approach is to 'divide and conquer', seeking *distinct* analyses of what are *distinct kinds* of causal relation. (Hall [2003] defends this position.)

At any rate, there is an issue here that strikes much deeper than the question concerning which of the various extant accounts of causation is best, and it is this: there appears to be a deep tension in our thinking about causation, which can be brought out by contrasting cases of late pre-emption and cases of causation by double prevention (Figure 9). Late pre-emption posed a problem that, we argued, may be solvable only if appeal is made to some principle to the effect that the causal structure of a process is intrinsic to it. Our example of double prevention appears to exhibit a cause *C* of an effect *E*—but, if so, it *contradicts* our intrinsicness thesis. What to do? It is attractive to suppose that causation is governed by some kind of intrinsicness thesis, as witness the most natural way of explaining our judgements about late pre-emption. It is attractive to suppose that there can be causation by double prevention. But on the face of it, we cannot hold both of these claims together, and that is a striking result. Should we simply soldier on, trying to find some analysis (most likely: a de facto dependence analysis) that can successfully navigate between Figures 6 and 9 (not to mention other examples we have not had the space to discuss)? Should we perhaps hold instead that the kind of causation that respects an intrinsicness thesis is not the same as the kind of causation exhibited in Figure 9? Should we give up in despair, concluding that our concept of causation is too much of a muddle to merit any kind of precise philosophical treatment?

We leave those questions with readers, hoping they will treat them with the seriousness they deserve. For we think that there is no hope of undertaking a

meaningful philosophical investigation of causation without addressing them. The days of seeking out clever counterexamples—while ignoring the deeper issues that lie behind them—are over.

References

Anscombe, G. E. M. [1971]: *Causality and Determination: An Inaugural Lecture*, Cambridge: Cambridge University Press.

Bennett, J. [1988]: *Events and their Names*, Indianapolis: Hackett.

Cartwright, N. [1983]: *How the Laws of Physics Lie*, Oxford: Clarendon Press.

Cartwright, N. [1999]: *The Dappled World*, Oxford: Oxford University Press.

Collins, J., Hall, N., and Paul, L. A. (eds.) [2003*a*]: *Causation and Counterfactuals*, Cambridge, MA: MIT Press.

Collins, J., Hall, N., and Paul, L. A. [2003*b*]: 'Counterfactuals and Causation: History, Background and Prospects', in Collins *et al.* [2003*a*].

Davidson, D. [1967]: 'Causal Relations', *Journal of Philosophy*, **64**, pp. 691–703.

Davidson, D. [1969]: 'The Individuation of Events', in N. Rescher (ed.), *Essays in Honor of Carl G. Hempel*, Dordrecht: Reidel, pp. 216–34.

Davidson, D. [1970]: 'Events as Particulars', *Noûs*, **4**, pp. 25–32.

Dowe, P. [2000]: *Physical Causation*, New York: Cambridge University Press.

Eells, E. [1991]: *Probabilistic Causality*, Cambridge: Cambridge University Press.

Ehring, D. [1997]: *Causation and Persistence*, New York: Oxford University Press.

Elga, A. [2000]: 'Statistical Mechanics and the Asymmetry of Counterfactual Dependence', *Philosophy of Science*, **68** (suppl.), pp. 313–24.

Fair, D. [1979]: 'Causation and the Flow of Energy', *Erkenntnis*, **14**, pp. 219–50.

Ganeri, J., Noordhof, P., and Ramachandran, M. [1996]: 'Counterfactuals and Preemptive Causation', *Analysis*, **56**, pp. 219–25.

Ganeri, J., Noordhof, P., and Ramachandran, M. [1998]: 'For a (Revised) PCA-analysis', *Analysis*, **58**, pp. 45–7.

Hall, N. [2000]: 'Causation and the Price of Transitivity', *Journal of Philosophy* **97**, pp. 198–222; reprinted in Collins *et al.* [2003*a*].

Hall, N. [2002]: 'The Intrinsic Character of Causation', in Zimmerman [2003].

Hall, N. [2003]: 'Two Concepts of Causation', in Collins *et al.* [2003*a*].

Hall, N., and Paul, L. A. [2002]: 'Causation and the Counterexamples: A Traveler's Guide', unpublished paper.

Hitchcock, C. [2001]: 'The Intransitivity of Causation Revealed in Equations and Graphs', *Journal of Philosophy*, **98**, pp. 273–99.

Hitchcock, C. [2003]: 'Do All and Only Causes Raise the Probabilities of Effects?', in Collins *et al.* [2003*a*].

Kim, J. 1973. 'Causation, Nomic Subsumption, and the Concept of Event', *Journal of Philosophy*, **70**, pp. 217–36.

Kvart, I. [2003]: 'Causation: Probabilistic and Counterfactual Analyses', in Collins *et al.* [2003*a*].

Lewis, D. [1973]: 'Causation', *Journal of Philosophy*, **70**, pp. 556–67; reprinted in Lewis [1986*a*], pp. 159–72.

Lewis, D. [1979]: 'Counterfactual Dependence and Time's Arrow', *Noûs*, **13**, pp. 455–76; reprinted with Postscripts in Lewis [1986*a*], pp. 32–66.

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

Lewis, D. [1986*b*] 'Postscripts to "Causation" ', in Lewis [1986*a*], pp. 172–213.

Lewis, D. [1986*c*]: 'Events', in Lewis [1986*a*], pp. 241–69.

Lewis, D. [2000]: 'Causation as Influence', *Journal of Philosophy*, **97**, pp. 182–97.

Lewis, D. [2003]: 'Causation as Influence', in Collins *et al.* [2003*a*]. This is an expanded version of Lewis [2000].

Lombard, L. [1986]: *Events: A Metaphysical Study*, London: Routledge & Kegan Paul.

McDermott, M. [1995]: 'Redundant Causation', *British Journal for the Philosophy of Science*, **46**, pp. 523–44.

Mackie, J. L. [1965]: 'Causes and Conditions', *American Philosophical Quarterly*, **2**, pp. 245–64.

Maudlin, T. [2003]: 'Causes, Counterfactuals and The third Factor', in Collins *et al.* [2003a].

Mellor, D. H. [2003]: 'For Facts as Causes and Effects', in Collins *et al.* [2003*a*].

Menzies, P. [1996]: 'Probabilistic Causation and the Pre-Emption Problem', *Mind*, **105**, pp. 85–117.

Noordhof, P. [1998]: 'Problems for the M-Set Analysis of Causation', *Mind*, **107**, pp. 457–63.

Paul, L. A. [1998*a*]: 'Problems with Late Preemption', *Analysis*, **58**, pp. 48–53.

Paul, L.A. [1998*b*]: 'Keeping Track of the Time: Emending the Counterfactual Analysis of Causation', *Analysis*, **58**, pp. 191–8.

Paul, L. A. [2000]: 'Aspect Causation', *Journal of Philosophy*, **97**, pp. 235–56; reprinted in Collins *et al.* [2003*a*].

Pearl, J. [2000]: *Causality: Models, Reasoning and Inference*, Cambridge: Cambridge University Press.

Ramachandran, M. [1997]: 'A Counterfactual Analysis of Causation', *Mind*, **151**, pp. 263–77.

Ramachandran, M. [1998]: 'The M-set Analysis of Causation: Objections and Responses', *Mind*, **107**, pp. 465–71.

Ramachandran, M. [2003]: 'A Counterfactual Analysis of Indeterministic Causation', in Collins *et al.* [2003*a*].

Salmon, W. [1994]: 'Causality Without Counterfactuals', *Philosophy of Science*, **61**, pp. 297–312.

Schaffer, J. [2000*a*]: 'Overlappings: Probability-Raising without Causation', *Australasian Journal of Philosophy*, **78**, pp. 40–6.

Schaffer, J. [2000*b*]: 'Trumping Preemption', *Journal of Philosophy*, **97**, pp. 165–81; reprinted in Collins *et al.* [2003*a*].

Schaffer, J. [2000*c*]: 'Causation by Disconnection', *Philosophy of Science*, **67**, pp. 285–300.

Schaffer, J. [2001]: 'Causation, Influence, and Effluence', *Analysis*, **61**, pp. 11–19.

Tooley, M. [1990]: 'Causation: Reductionism versus Realism', *Philosophy and Phenomenological Research*, **50** (suppl.), pp. 215–36.

van Inwagen, P. [1978]: 'Ability and Responsibility', *Philosophical Review*, **87**, pp. 201–24.

Yablo, S. [1992]: 'Cause and Essence', *Synthese*, **93**, pp. 403–49.

Yablo, S. [2003]: 'Advertisement for a Sketch of an Outline of a Proto-Theory of Causation', in Collins *et al.* [2003*a*].

Zimmerman, D. [2003]: *Oxford Studies in Metaphysics*, Vol. 1, Oxford: Oxford University Press.