JOHN COLLINS

COUNTERFACTUALS, CAUSATION, AND PREEMPTION

1 INTRODUCTION

A *counterfactual* is a conditional statement in the subjunctive mood. For example:

> If Suzy hadn't thrown the rock, then the bottle wouldn't have shattered.

The philosophical importance of counterfactuals stems from the fact that they seem to be closely connected to the concept of *causation*. Thus it seems that the truth of the above conditional is just what is required for Suzy's throw to count as a cause of the bottle's shattering. If philosophers were reluctant to exploit this idea prior to 1970, it was because of a widespread feeling that the truth-conditions of the counterfactual conditional were not sufficiently well understood. The development of a formal semantics for counterfactuals by Robert Stalnaker [1968] and David Lewis [1973b] stands as a major recent achievement in philosophical logic.

Section 2 presents the standard Stalnaker–Lewis semantics for the counterfactual conditional and develops some of the logical features of counterfactuals. Section 3 presents Lewis's original counterfactual theory of causation [1973a], and explains the problems that eventually led him to abandon the theory in its original form. The remainder of the article surveys the current state of counterfactual theories of causation, by presenting, in Sections 4, 5, and 6, three recent contending accounts, due to Lewis [2000; 2004a], Yablo [2000; 2004] and Hall [2004b; 2004a].

The discussion does not aim to be exhaustive, but focuses on the central issue of *preemption*, which has proved to be the major hurdle for counterfactual theories. For a more complete picture, the reader is referred to the papers in the collection edited by Collins, Hall, and Paul [2004]. The present article aims to provide an brief introduction to this currently very lively area of applied philosophical logic. For a more comprehensive introductory survey, Hall and Paul (forthcoming) is also highly recommended.

The contemporary literature on counterfactuals and causation includes a vast and potentially bewildering collection of examples and counterexamples. The present article focuses on six central types of example, which are labeled (E1)–(E6) so that the reader may more easily distinguish them from other less important particular cases mentioned in passing.

2 THE LOGIC OF COUNTERFACTUALS

The counterfactual 'If A were true, then C would be true' with antecedent A and consequent C is sometimes written '$A\Box \rightarrow C$' in order to distinguish it from other kinds of conditional statement. For example the counterfactual must be distinguished from both the material conditional of first-order logic, and the "strict conditional" of entailment. The truth-functional material conditional '$A \rightarrow C$' is logically equivalent to '$\sim A \vee C$' and thus has truth-conditions weaker than those of the corresponding counterfactual conditional. Not every counterfactual with a false antecedent or a true consequent is true. The strict conditional, on the other hand, is too strong; it might be true that:

> If this match had been struck, it would have lighted.

But the lighting of the match is not logically entailed by its being struck.

A correct account of the semantics of the counterfactual will then, presumably, locate it somewhere between these two extremes. But where? One obvious thought is that '$A\Box \rightarrow C$' is true if and only if C is entailed not by A alone, but by A in conjunction with certain other truths, including, perhaps, the laws of nature. Thus it might well be that the lighting of the match *is* entailed by its being struck, in conjunction with the laws of nature, and certain other true matters of fact — for example the presence of sufficient oxygen in the atmosphere.

The problem with this idea is that there is no single fixed set of truths that will do the job for all A and C. That is because counterfactuals are *non-monotonic*. In other words the inference pattern:

$$\begin{array}{ll} & \underline{A\Box \rightarrow C} \\ \text{so:} & (A \,\&\, B)\Box \rightarrow C \end{array}$$

is invalid. It may be true, for example, that the match would light if struck, and yet not true that it would light if struck *in the absence of oxygen*.

But the counterfactual conditional will only serve as a fit tool for philosophical analysis if we have a firm grasp of its logic and truth-conditions. In a famous essay critical of philosophical use of the counterfactual idiom, Nelson Goodman framed the challenge this way. In evaluating the truth of the counterfactual '$A\Box \rightarrow C$' we want to hold fixed all those truths that are "cotenable" with the truth of the antecedent A. Yet what might it mean for a proposition to be cotenable with A other than that the proposition is one that would still be true even if A were true? If one hopes to provide a semantics for counterfactuals in terms of what is cotenable with a given antecedent, then one had better not rest this on a counterfactual analysis of cotenability [Goodman, 1947].

It was the work of Stalnaker [1968] and Lewis [1973b], which in turn grew out of the development of possible world semantics for modal logic in the 1960s, that first convinced some of the skeptics that counterfactual conditionals were indeed philosophically respectable.

The Stalnaker–Lewis approach starts from the assumption that the set of all possible worlds may be weakly ordered with respect to the "comparative similarity" of those world to the actual world. Since this ordering is transitive and connected, it is useful heuristically to think of it as a comparative "closeness" relation. If A is any proposition, call a world at which A is true an *A-world*. Then, in Lewis's formulation, the truth-condition for the counterfactual may be stated in this way:

> '$A\Box\rightarrow C$' is true if and only if some $(A\ \&\ C)$-world is more similar to the actual world than any $(A\ \&\ \sim C)$-world is.

If we simplify things by assuming that for each A there is always a single A-world most similar to the actual world, then the condition becomes:

> '$A\Box\rightarrow C$' is true if and only if C is true at the most similar A-world to the actual world.

Now similarity is itself a philosophically problematic concept, but even without settling on any particular criteria for making comparative judgments, we can see how much of the logic of counterfactuals follows simply from the fact that any candidate similarity ordering, like a closeness ordering, must be connected and transitive.

We can see, for example, why strengthening the antecedent of a counterfactual may lead from truth to falsehood. The fallacy is akin to that made a person who infers from the fact that there is no bank in the closest town to here, that there is no bank in the closest town to here with a restaurant.

We can also see immediately that contraposition fails for the counterfactual conditional, i.e., that the inference

$$\begin{array}{ll} & \underline{A\Box\rightarrow C\quad\ } \\ \text{so:} & \sim C\Box\rightarrow\sim A \end{array}$$

is fallacious. It doesn't follow from the fact that the closest town with a restaurant has no bank that the closest town with a bank has no restaurant.

A similar exploitation of the analogy with spatial closeness will convince us that counterfactuals are not transitive. That is, we can add to our list of counterfactual fallacies the following:

$$\begin{array}{ll} & A\Box\rightarrow B \\ & \underline{B\Box\rightarrow C} \\ \text{so:} & A\Box\rightarrow C. \end{array}$$

Moving beyond these purely logical points, however, more will need to be said about the comparative similarity relation. The problem is that similarity admits of different respects. One possible state of affairs may be more similar to actuality than another in one respect, and less similar to actuality in another. Often it is the context of utterance that determines the particular similarity ordering that a

speaker has in mind. But insofar as we want counterfactuals to provide objective truth-conditions for causal statements, this question of context-dependence is fairly pressing. How are the various respects of similarity to be weighed in order to enable a single overall comparison?

It seems clear that similarities with respect to the laws of nature should generally outweigh similarities with respect to accidental matters of fact. But this cannot be an invariable rule. For if, as we are assuming, the laws of nature are deterministic in both temporal directions, any supposed change in the way things presently are, will be propagated, via the laws, into a divergent past as well as a divergent future. Matching the past history of the actual world is very important for similarity, it appears, even if the match can be obtained only by allowing a minor localized violation of the laws of nature (a "small miracle").

This is not to deny that we *are* sometimes prepared to speak as though things would have been different in the past had they been different now. For example I might say: "If I had jumped out the window just now, there would have to have been a safety net in place, I'm not crazy!" But note the "have to have been" construction in that sentence that serves as a syntactic indicator of the appropriateness of the *backtracking* interpretation (see [Lewis, 1979]). The point is simply that this backtracking interpretation is non-standard. Causes always, or at least typically, precede their effects. Surely this asymmetry should be reflected by a corresponding asymmetry in the counterfactuals.

But since the temporal asymmetry of causation seems to be a merely contingent feature of the actual world, it seems wrong to build this asymmetry into the analysis of counterfactuals by stipulation. Lewis [1979] pursues the more ambitious goal of identifying criteria for a comparative similarity relation that rule out backtracking in worlds like the actual world, but without making backward causation an *a priori* impossibility.

Here is the rough idea. Suppose that c is some event that actually occurred at time *t*, and consider a counterfactual whose antecedent asks us to suppose that c hadn't occurred. A non-backtracking reading of this counterfactual will be ensured provided that all the closest possible worlds to the actual world at which c doesn't occur are worlds in which (i) past history up until some point shortly before time *t* perfectly matches the history of the actual world, and (ii) this perfect match results from a small "divergence" miracle. And this is correct, Lewis suggests, because our world happens to be such that there is no possible world *w* in which (i) c fails to occur, (ii) the future of *w after* time *t* exactly matches the actual future, and (iii) this match results from a small "reconvergence" miracle.

An important recent paper by Adam Elga has raised a serious difficulty for Lewis's attempt to rule out backtracking contingently. Elga argues that statistical mechanics provides examples which demonstrate that reconvergence to the actual world requires no greater violation of the laws than divergence from it does. See [Elga, 2000; Albert and Loewer, 2005].

3 THE COUNTERFACTUAL THEORY OF CAUSATION

An early statement of a counterfactual analysis of causation can be found in the work of David Hume [1902, § VII], where he writes:

> ...we may define a cause to be *an object followed by another, and where all the objects, similar to the first are followed by objects similar to the second.* Or, in other words, *where, if the first object had not been, the second never had existed.*

In fact, of course, the passage just quoted contains two quite different accounts of the causal relation. While the second sentence treats causation in counterfactual terms, the first expresses an idea that is closer to what has become known as a *regularity theory* of causation.

Regularity theories tended to be the more favored of these two broad approaches prior to the development of the Stalnaker–Lewis semantics for the counterfactual. See, e.g., [Mackie, 1965].

A regularity analysis of causation states, roughly, that:

> One event c is a cause of another event E just in case c is a member of some minimal set of actual events that are jointly sufficient, given the laws of nature, for the occurrence of the effect.

But regularity theories of causation have always faced difficulties. One is the problem of distinguishing cause from effect. For if E is caused by c, there will often be a set of conditions including e which, in conjunction with the laws, entail that c occurs.

Another difficulty is the problem of distinguishing genuine causes from inefficacious epiphenomenal by-products of a causal process. If d and e are independent effects of some common cause c, then it may well be the case that d belongs to a minimal set of conditions which, along with the laws, are sufficient for e. Then d, which is a epiphenomenal by-product of the process by which c caused e, will be incorrectly counted as a cause of e.

In his seminal article [1973a], Lewis pointed out that a counterfactual analysis of the causal relation is at a distinct advantage over a regularity theory in both of these cases. Given an appropriate ban on a backtracking interpretation of the relevant counterfactuals, one may simply deny the truth of the counterfactuals that might otherwise cause problems. Suppose that e wouldn't have occurred if c hadn't occurred. Might it also be the case that if e hadn't occurred then c wouldn't have? Not unless the latter conditional is understood in the backtracking sense. Similarly, if *d* and e are independent effects of a common cause c, then the only reason one might be tempted to believe that *if d hadn't occurred then e wouldn't have occurred* is if one thinks that *if d hadn't occurred then neither would c*. Once again, this would involve an illicit backtracking reading of the latter conditional.

Say that one event e is *counterfactually dependent* on another event c when e wouldn't have occurred if c hadn't. The simplest counterfactual analysis of causation would simply construe causation as counterfactual dependence.

But this simplest account cannot be correct, as is demonstrated by the following example:

(E1) *Early Preemption*: Suzy throws a rock at a bottle. The rock hits the bottle and shatters it. Billy was standing by with a second rock. Had Suzy not thrown her rock, Billy would have shattered the bottle by throwing his rock.

The problem here is that the shattering of the bottle is indeed caused by Suzy's throw, despite the fact that the shattering is not counterfactually dependent on the throw. According to Lewis, the key to reinstating the throw as cause is to recognize that there is a *chain of events* initiated by the throw and terminating in the shattering, such that each event in the chain is counterfactually dependent on the one before it. Call such a chain of events a *chain of counterfactual dependence*. In this case, each event in the chain is the event of Suzy's rock being located in midflight at some point between her hand and the bottle. Although Billy was standing by with appropriate intent and deadly aim, no such chain of events connects his standing by with the shattering.

In cases like Early Preemption we have a *chain of counterfactual dependence* leading from a cause to an effect, without the effect being counterfactually dependent on the cause. The problem with preemption, then, seems to stem directly something we noted in the previous section: the non-transitivity of the counterfactual conditional. While causation appears to be a transitive relation, counterfactual dependence is not.

If this diagnosis of the problem is correct, then the fix is very straightforward. We should simply take causation to be the *ancestral* of the counterfactual dependence relation. Thus we arrive at Lewis's [1973a] counterfactual analysis of causation.

(C1) Suppose that c and e are distinct events, and let C and E respectively be the propositions that the events c and e occur. Then e is *counterfactually dependent* on c when the following two counterfactual conditionals are true:

(i) $C\Box\!\!\rightarrow E$

(ii) $\sim C\Box\!\!\rightarrow \ \sim E$

(C2) A *causal chain* is a finite sequence of actual events such that each event in the sequence is counterfactually dependent on the previous event.

(C3) One event is a *cause* of another if and only if there is a causal chain leading from the first to the second.

Note that it is precisely the failure of contraposition for the counterfactual conditional that leaves room for the ban on backtracking, by allowing that e may be

counterfactually dependent on c without c also being counterfactually dependent on e.

This original analysis was designed to deal with cases of preemption like (E1). However the phenomenon of preemption has proved to be a far more serious difficulty for the counterfactual theory of causation than Lewis believed it to be in 1973.

The problem is that there are varieties of preemption that cannot be dealt with by the transitivity strategy. Here is one such:

(E2) *Late Preemption*: Billy and Suzy both throw rocks at a bottle. Suzy's rock gets there first, hitting the bottle and shattering it. Billy's rock flies through the now empty space where the bottle was standing.

This example differs in one key respect from the previous case. In Late Preemption no stepwise dependent chain of events can be traced from Suzy's throw to the shattering of the bottle. For consider all the events in the process initiated by Suzy's throw prior to the shattering, that is: Suzy's rock being located at various positions at various times in mid-flight. The shattering fails to depend on *any* such event, because, had Suzy's rock failed to be there, the bottle would still have been shattered by Billy's rock.

Here is another way of seeing the difference between the Early and Late Preemption examples. In cases like Early Preemption, the possible process initiated by the preempted standby that would otherwise have led to the effect, is cut off by the process leading from the preempting cause to the effect, at some time *before* the effect occurs. In Late Preemption, this "cutting-off " takes place only when the effect occurs. (This way of looking at things explains the "early/late" terminology.)

Furthermore, as Jonathan Schaffer discovered, there are possible cases of preemption that do not involve any kind of cutting-off at all. In Schaffer's example of "Trumping Preemption" the absence of any cutting, either early or late, is guaranteed simply by stipulating that the causal process in question works by action at a distance.

(E3) *Trumping Preemption*: The laws of magic are such that what happens at midnight is determined by the first spell cast the previous day. At noon Merlin casts the first spell of the day: a spell to turn the Prince into a frog. At six that evening Morgana casts a spell to turn the Prince into a frog. At midnight the Prince turns into a frog.

This example also causes problems for Lewis's 1973 theory. The transfiguration of the prince is not counterfactually dependent on Merlin's spell, since if Merlin had not cast the spell, the prince would still have been turned into a frog by Morgana's spell. In addition, there is no chain of counterfactual dependence leading from Merlin's spell to the transfiguration, since the example stipulates that none is required.

It is tempting to dismiss such fantastic cases as being too far-fetched to be relevant to any discussion of causation as it is in the actual world. But this would be

too hasty a dismissal. The fanciful nature of Schaffer's story merely helps make the causal structure of the example clear. The important point is that it seems perfectly possible that, for all we know, some actual cases of causation work that way. Since the possibility of causation by trumping preemption cannot be ruled out *a priori*, a theory of causation will only be adequate if it can deal with such cases.

4 CAUSATION AS INFLUENCE

One tempting thought is that the problems facing the counterfactual theory might be solved by taking events to be modally "fragile", i.e., by claiming that any difference in time, or manner of occurrence makes for a numerically distinct event. In our Late Preemption example, the bottle would still have been shattered had Suzy not thrown her rock, but it would have been shattered a moment later by Billy's rock, and, presumably, shattered in a slightly different way. Thus, if the actually occurring shattering is construed as a modally fragile event (i.e., an event with a particularly rich essence) then the counterfactual dependence of the actual shattering on Suzy's throw is restored, since, had Suzy's throw not taken place, the shattering that would have occurred instead would have been a *different* shattering; a numerically distinct event.

This "fragility strategy" is discussed by Lewis in various places (e.g., [1986b, pp. 193–199] and [2004a, pp. 85–90]. As Lewis recognizes, there are various reasons for rejecting the approach.

For one thing, the "uncommonly stringent" conditions of occurrence that the fragility strategy imposes are at odds with our ordinary standards of event identity. Lewis points out that we are usually quite happy to allow that one and the same event might have been delayed, as, for example, when a seminar talk is postponed rather than cancelled [Lewis, 2004a, p. 86].

Secondly, the fragility strategy produces spurious cases of causation, as in Lewis's example of the "Poison and the Pudding". Suppose a poison kills its victim more slowly and painfully when taken on a full stomach. The victim eats some pudding and then drinks the poison. If the victim's actual death is construed as modally fragile, then it is an event that would not have occurred had the pudding not been eaten. Yet the eating of the pudding was not a cause of his death [Lewis, 1986b, pp. 198–199].

Still, the central idea of fragility has a role to play in Lewis's [2000; 2004a] revised account of "causation as influence". While remaining neutral about the ordinary identity conditions for events, Lewis proposes that we introduce a new technical term to refer to events *construed as* modally fragile. Say that an *alteration* of an event is either the very fragile version of the event that actually occurs, or a fragile alternative to it that differs slightly with respect to time or manner of occurrence. Lewis then suggests that we "look at the pattern of counterfactual dependence of alterations of the effect upon alterations of the cause". Say that event c *influences* event e when there is a substantial range c_1, c_2, c_3, ... of alterations

of c, and a substantial range e_1, e_2, e_3, ... of alterations of *e*, such that if c_1 had occurred then e_1 would have occurred, and if c_2 had occurred then e_2 would have occurred, and so on [Lewis, 2004a, p. 91].

Our original notion of counterfactual dependence was a notion of "whether-whether" dependence. One event is dependent on another is this sense just in case whether or not the event occurs depends on whether or not the other occurs. But there are other varieties of dependence. Lewis's idea is that we should think of degree of causal influence as being determined by the extent to which *whether*, *when*, and *how* one thing happens depends counterfactually on whether, when, and how, something else happens. Hall and Paul have usefully labeled this idea the "counterfactual covariation" theory of causation (Hall and Paul, forthcoming).

As in Lewis's original counterfactual theory, causation is now taken to be the ancestral of the influence relation. One event c causes another event e if and only if there is a chain of stepwise influence from c to e.

Let's see how this idea works in our case of Late Preemption. If we consider small alterations to the time of Suzy's throw, we obtain corresponding alterations to the time of shattering (provided of course that the throw is not so much delayed that Billy's rock gets there first). Or consider small alterations to the manner in which Suzy throws her rock; alterations, perhaps, to the velocity and direction of her throw. Throughout a range in which the velocity is still great enough for her rock to beat Billy's to the bottle, and Suzy's aim is still accurate enough to score a fairly direct hit, alterations to the throw will produce a counterfactually covarying range of alterations to the shattering. So the shattering is influenced by Suzy's throw.

Not so for Billy's throw. Unless we imagine Billy's throw sufficiently altered in time or manner so that his rock reaches the target before Suzy's does, the extent to which the time and manner of the shattering depends on alterations to the time and manner of the throw are completely negligible. This, claims Lewis, is why Suzy's throw counts as a cause of the shattering and Billy's does not.

Lewis also suggests that the analysis of causation as influence provides a solution to the problem of Trumping Preemption. Although there is no influence of the whether-whether or when-when variety, the transformation of the prince is nevertheless influenced by Merlin's spell since the manner of transformation covaries with the kind of spell cast. For example: had Merlin uttered "Presto! Prince-to-possum!" instead of "Presto! Prince-to-frog!" at noon, then the prince would have turned into a possum, rather than a frog, at midnight. On the other hand, what happened at midnight was in no way dependent on whether, when, or how Morgana acted in the late afternoon.

However this solution to the Trumping problem seems to turn on inessential and eliminable features of Schaffer's original example. If we suppose that Merlin's options are limited to a single spell that he may cast (standard prince-to-frog) and a single time of day he may cast it (noon), then the transfiguration of the prince is no longer influenced, in Lewis's sense, by Merlin's spell, though of course Merlin's spell is still its cause. (See [Collins, 2000], in [Collins *et al.*, 2004, p. 114]. The

idea is due to Jacob Rosen.) This is a counterexample to the necessity of the influence theory. Other counterexamples to the necessity of the influence theory can be constructed from cases of Early or Late Preemption by ensuring that the preempted backup would have brought about the effect in exactly the same manner and at exactly the same time. See Strevens [2003] and Yablo's "Smart Rock" example reported in [Hall, 2004b, p. 237].

Schaffer has argued that the idea behind our first counterexample to necessity can also be developed into a counterexample to sufficiency. Add to the story of Merlin's limited options the fact that Morgana has a vast range of possible spells to cast, and available times at which to cast them. And now suppose Morgana stands by, silently watching, just before noon as Merlin prepares to cast his spell. Then, claims Schaffer, the transfiguration at midnight is influenced by Morgana's silent watching (given her vast range of options) though her watching is not among its causes. The point here seems less clear than in the previous case, since Schaffer must counter the suggestion that Morgana's watching *is* a cause — by *omission* — of the prince's becoming a frog. For more details, see [Schaffer, 2001]. Collins suggests that Lewis's own "Poison and Pudding" example was already a counterexample to the sufficiency of the causation as influence account [2000].

5 *DE FACTO* DEPENDENCE

In our Early Preemption example, the shattering of the bottle was not counterfactually dependent on Suzy's throw, which, nevertheless, caused it. Yet note this: holding fixed the fact that Billy does not throw his rock, the shattering *does* depend on Suzy's throw. That is, if Suzy hadn't thrown her rock and Billy had still not thrown his rock either, then the bottle would not have shattered. Stephen Yablo [2000; 2004] suggests we say that the shattering has a "*de facto* dependence" on Suzy's throw, and count Suzy's throw as a cause of the shattering in virtue of this *de facto* dependence. A version of Yablo's proposal will be developed in this section. Closely related accounts have been proposed by Judea Pearl [2000] and Christopher Hitchcock [2001].

In Late Preemption the same idea seems to work. Holding fixed the fact that Billy's rock doesn't hit the bottle, the shattering depends on Suzy's throw. That is, if Suzy hadn't thrown her rock, and yet Billy's rock still hadn't hit the bottle, then the bottle would not have shattered. Admittedly, this last counterfactual is a little weird. Since no backtracking is allowed, and we are holding fixed the fact that Billy's rock doesn't hit the bottle, the antecedent is asking us to suppose that Billy's rock is thrown just as it actually was, that it follows the same deadly-accurate trajectory toward the bottle (which is still there when it arrives, since Suzy hasn't thrown her rock) but then, somehow — miraculously! — fails to hit it.

Clearly some restrictions will have to be placed on the kind of proposition that may be held fixed, otherwise everything will turn out to depend on everything else given some suitably cooked-up background condition. For consider any simple

and unproblematic case of causation: Billy acting alone throws a rock that shatters a bottle. Let B be the proposition that Billy throws the rock. Now let L be any other true proposition. L may be completely unrelated to, and independent of, the process that leads to the shattering of the bottle, for example the proposition that far away from the scene of the bottle breaking a certain leaf flutters down to the ground from a tree. Now let G be the proposition: B or not-L. Holding this proposition G fixed, the shattering of the bottle depends on the fall of the leaf.

In fact, for any pair of actual events whatsoever, we can simply let G be the proposition that either both of the events occur or neither does. Holding this proposition G fixed, each of the events depends on the other.

Allowing the choice of these kinds background conditions would trivialize the *de facto* dependence proposal. Yablo proposes to bar such choices on the grounds that are not sufficiently "natural", but he does not fully explain what naturalness amounts to in this context. The two examples just given suggest that it may be the disjunctiveness of the proposition in question that renders it unnatural, and that such background conditions are to be barred for the same reasons that disjunctive events are disallowed on many theories of events.

Even with this restriction, however, the proposal as sketched so far founders on cases of the following kind (due originally to Hartry Field):

(E4) *Self-Countering Threat*: Billy places a bomb under Suzy's desk. Suzy notices the bomb and takes cover before it explodes. Suzy survives uninjured.

Holding fixed the true, and perfectly natural, proposition that there was an explosion under Suzy's desk, Suzy's survival *de facto* depends on Billy's planting of the bomb. That is, if Billy hadn't planted the bomb, and yet there had still been an explosion under her desk, then the bomb would not have been there for Suzy to notice, so she would not have taken cover, and hence would not have survived uninjured. Again, one should not be put off by the fact that this last counterfactual shares the strangeness of the corresponding counterfactual in our discussion of the Late Preemption example. Since no backtracking is allowed, and we are holding fixed the fact that there is an explosion under Suzy's desk, the antecedent is asking us to suppose that although there is nothing under Suzy's desk to lead her to take cover, there is, nevertheless, an explosion there.

Thus the proposal developed so far rules Billy's planting of the bomb to be a cause of Suzy's survival. But this seems quite incorrect, for it was Billy's action that threatened Suzy's survival in the first place. Billy's action can be thought of as a *self-countering threat* because it both threatens to prevent something, and also prevents that threat from succeeding. Intuitively: self-countering threats to an event's occurrence don't count as causes of it.

Yablo's strategy in the face of this problem is to attempt a principled explanation of the sense in which the dependence in these cases is "artificial".

His proposal is this: Suppose that E depends on C given H. Then say that the dependence of E on C given H is *artificial* if and only if every proposition that E depends on given that *not-*C is also something that E depends on given H. As

Yablo puts it, the idea here is that the dependence on C is artificial just in case C "addresses some need" that is simply added to what would, if C had been false, have been all the needs.

Let's see how this works in our example of Self-Countering Threat. We want the dependence of Suzy's survival on the bomb planting, given the explosion, to turn out to be artificial. Is it? Well consider all those things that Suzy's survival would have depended on if Billy hadn't placed the bomb under her desk. Remember: we are now *not* holding fixed the fact of the explosion. In this "fallback scenario" as Yablo calls it, Suzy's survival depends on the usual mundane sorts of thing that all of us require in order to get through the day intact. It depends, for example, on her heart and lungs continuing to function normally, on her not being struck by lightning, and so on and so forth. Now the point is that all of these quite mundane things on which Suzy's survival depends in the fallback scenario, are also things on which her survival depends given that there is an explosion under her desk. No use avoiding the bomb blast and then suffering a fatal heart attack. It follows that the dependence of Suzy's survival on the planting of the bomb (holding the explosion under the desk fixed) is an artificial dependence.

But of course this fix is useless if it robs us of what we thought we had already obtained: a solution to the problem of Late Preemption. Returning to that example we must check that the dependence of the shattering on Suzy's throw, given that Billy's rock does not hit the bottle, does not also turn out to be artificial.

This dependence will prove to be artificial only if every event on which the shattering would have depended had Suzy not thrown her rock, is also an event that the shattering depends on given that Billy's rock doesn't hit the bottle. Now suppose that Suzy had not thrown her rock. Then the bottle would still have been shattered by Billy's rock, and the shattering would have depended on Billy's throw. But is Billy's throw also an event on which the shattering depends, given that Billy's rock does not hit the bottle? An odd question, perhaps, but one that is not difficult to answer in the negative if we keep the meaning of the "given that" locution firmly in mind. For clearly, if Billy had not thrown his rock, and Billy's rock had not hit the bottle, then the bottle would still have been shattered (by Suzy's) rock. So the *de facto* dependence of the shattering on Suzy's throw, given that Billy's rock does not hit the bottle, is not an artificial one.

Now anything can be made to look dependent on anything else, if we are allowed to hold fixed background conditions that are not sufficiently natural. That a similar trick might be used to show that every dependence is artificial, is the thought behind the framing of clause (D3) in the following formulation of Yablo's *de facto* dependence account of causation:

(D1) Say that *E depends on C given G* when the following counterfactual is true: If C were false and yet G were still true, then E would be false.

(D2) Say that the dependence of E on C given G is *artificial* when every proposition on which E would depend if C were false is also a proposition on which E depends given G.

(D3) Say that the proposition E de facto *depends* on C if and only if there is some true proposition G, such that E depends non-artificially on C given G, and such that G is more natural than any proposition *H*, given which the dependence of E on C is artificial.

Finally the proposal is that:

(D4) One event c is a *cause* of another event e if and only if the proposition that e occurs *de facto* depends on the proposition that c occurs.

One advantage that Yablo claims, if very tentatively, for the *de facto* dependence account, is that, unlike other approaches, it is "not at an absolute loss" when it comes to dealing with the Trumping Preemption examples [2004, p. 134].

The idea here is that we can demonstrate the *de facto* dependence of the prince's transformation on Merlin's spell by holding fixed the proposition that no-one casts a spell before Merlin does. If Merlin had cast no spell then, given that no-one casts a spell before Merlin does, the prince wouldn't have turned into a frog.

Is this dependence artificial? Only if every event on which the transformation would have depended had Merlin cast no spell, is also an event the transformation depends on given that no-one casts a spell before Merlin. Is this true? Well, if Merlin had cast no spell, the prince's transformation would have depended on Morgana's six o'clock spell. But given that no-one casts a spell before Merlin, does the transformation depend on Morgana? No, because if Morgana had cast no spell, then Merlin would still have cast his (this being compatible with the proposition being held fixed) and the prince would still have turned into a frog.

But the real problem with this proposal for dealing with Trumping Preemption on the *de facto* dependence account seems to be with the degree of naturalness of the proposed background condition, rather than with the artificiality of the resulting dependence. To say that *no-one casts a spell before Merlin* is simply equivalent to saying that *either Merlin casts the first spell, or there is no spell cast at all*. The proposed condition is exposed as disjunctive!

The *de facto* dependence faces its own difficulties as well. It is not clear, for one thing, that Yablo has really succeeded in making clear the "E depends on C given G" locution; recall those weird counterfactuals that arose in the discussion of the examples. The same difficulty, it should be noted, also faces Pearl's and Hitchcock's version of the theory, though in those cases the problem is made less visible, since a solution to it is simply assumed to have been built into the framework in which a particular causal situation is to be modeled.

A second problem facing the theory has to do with the crucial notion of "naturalness", which Yablo has not fully succeeded in explaining.

Finally, Hall and Paul (forthcoming) have constructed a new and interesting case of preemption that appears to make trouble for the *de facto* dependence theory.

(E5) *Partial Preemption*: Billy and Suzy are together attempting to push a heavy box across the floor. Billy is stronger than Suzy. He could shift the box all

> by himself if he really wanted to. Suzy couldn't. But Billy is also lazier than Suzy. When he sees that Suzy is using all of her strength, Billy decides to exert less effort than he might, and he applies a force insufficient on its own to move the box. Nevertheless their combined efforts succeed in shifting it.

Suzy's push and Billy's push are, intuitively, joint causes of the shifting. Yet the shifting fails to depend on Suzy's push, since had Suzy not pushed, Billy would have pushed harder and moved the box by himself.

Now let G be the proposition that Billy believes that he can exert less effort than he might, since Suzy is using all of her strength. Given G, the shifting depends on Suzy's push. But Yablo's account must deem this dependence artificial, since every event on which the shifting would have depended had Suzy not pushed, is also an event on which the shifting depends given that Billy believes he can exert less effort than he might.

6 THE BLUEPRINT STRATEGY

A third recent proposal that shows some promise in addressing the problems posed by the various preemption cases is the "blueprint strategy" developed and defended by Ned Hall [2004a; 2004b].

The idea behind the blueprint strategy is that we should hold fast to the thought that causation is an intrinsic matter; that whether or not a process is a causal one ought to depend only on the intrinsic nature of the process itself, and on the laws of nature.

In some ways the blueprint strategy might be thought of as a more sophisticated descendent of Lewis's 1986 account of causation as "quasi-dependence" (see [1986b, pp. 193–212]) although, as we shall see, development of the strategy will turn out to require a quite radical break with the whole tradition of thinking of causation in terms of counterfactual dependence.

Here are the details of the quasi-dependence theory. Call a possible process that exactly matches all of the intrinsic properties of another process a *duplicate* of that process. If a possible process that is a duplicate of some process P exists in a possible world that has the same laws of nature as the world in which P is located, call it an *isonomic duplicate* of P. If causation is an intrinsic matter, then any isonomic duplicate of a causal process must also be a causal process. Suppose that some actual process is an isonomic duplicate of a possible process whose last event depends counterfactually on its first. Then say that the last event of the actual process is *quasi-dependent* on its first event. A *causal chain* is then taken to be any chain of events in which each event is either counterfactually dependent or quasi-dependent on the preceding event. One event is a *cause* of another if and only if there is a causal chain leading from the one to the other.

If this idea can be made to work, then it offers a ready solution to the problem of Late Preemption. The shattering of the bottle does not depend counterfactually

on Suzy's throw, because of the backup process initiated by Billy when threw his rock a moment later. Yet it is not hard to find an isonomic duplicate of the process leading from Suzy's throw to the shattering, for which the counterpart shattering depends counterfactually on the counterpart of Suzy's throw. Just consider a possible scenario that is exactly like the actual one except that Billy does not throw his rock. The quasi-dependence theory thus correctly counts our case of Late Preemption as causal. Hall seeks to procure the same kind of solution to the preemption problem. In the example just discussed, the process initiated by Suzy's counterpart that leads to the shattering of the bottle's counterpart is an example of what Hall calls a *causal blueprint* for the actual process it duplicates. The blueprint process is uncontroversially causal, and we then appeal to the thesis that causation is an intrinsic matter in order to establish that the problematic actual process is a causal one too.

It is important here that a *process* be understood as something more than just an arbitrary chain or sequence of events. For it is simply not true that any isonomic duplicate of a chain of counterfactual dependence is a causal chain, as can be seen from the following example:

Dominoes: A large number of dominoes, standing on end, are arranged in a line so that when I push the first domino over it strikes the second, which falls and topples the third, and so on. The fall of the last domino is counterfactually dependent on the fall of the first one. But now consider a similar setup in which all except the first two and the last two dominoes have been removed. I topple the first domino so that it hits the second, and then later push over the second last domino so that it hits the final one. We can imagine this done in such a way that the pair of events consisting of the fall of the second and the fall of the final domino is an isonomic duplicate of the corresponding pair of events in the original setup. Yet the first of these two events does not cause the second.

Examining this case, we can that it is crucial that the structure of events duplicated include all the causal intermediaries of any pair of events in the structure. Less obvious, perhaps, is the fact that, unless *all* the causes contributing to the effect are included, a causal process leading up to some event may have an isonomic duplicate that is not a cause of the counterpart event, as the following example shows:

Voting: Billy, Suzy, and their friend Biff are voting on a proposal. Their votes are submitted electronically and a light goes on as soon as the machine receives either two yes votes or two no votes. Scenario (A): Billy and Suzy simultaneously press their yes buttons and the light goes on. Biff abstains. Scenario (B): Billy, Suzy, and Biff all press voting buttons simultaneously. Suzy votes yes, Billy and Biff vote no. The light goes on. Suzy's pressing her yes button was a (joint) cause of the light's going on in (A) but not in (B). (Based on a schematic example in [Hall, 2004a, pp. 272–273].)

If we assume that there is neither backward causation nor causation at a temporal distance, this further condition may framed as the requirement that the structure of events to be duplicated consist of some event e along with *all* of the causes of e traced back to some earlier time.

One further adjustment needs to be made. The demand that the causal structures be perfect duplicates of one another is too strict. In our Late Preemption case, for example, the suggested blueprint process seems adequate despite the fact that it is a less than perfect intrinsic match of the original. But we are entitled to ignore the minute differences in the trajectory of Suzy's rock, due to subtle differences in the surrounding air currents, gravitational field and so on, that result from Billy's not throwing.

Putting all of this together, we arrive at Hall's formulation of the Intrinsicness thesis:

Intrinsicness: Suppose that S is a structure of events consisting of event e, together with all of its causes back to some earlier time. 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' [Hall, 2004a, p. 264].)

Of course, this is only part of what has to be done. We want to claim is that in problem cases like the preemption examples, causation amounts to duplication of a possible structure whose causal credentials are unproblematic (since the preempted backup processes are absent). But to complete the story, a defender of the blueprint strategy must explain what it is that constitutes the causal relation in the unproblematic situation in the first place. An obvious idea would be to appeal here to counterfactual dependence as a sufficient condition for causation in the blueprint, as Lewis hoped to do during the period in which he defended the quasi-dependence account. Then we could apply the Dependence thesis to establish that some blueprint is indeed a causal structure, and use the Intrinsicness thesis to carry this result over to solve the problematic preemption cases.

But unfortunately the obvious idea doesn't work, for the simple reason that the Intrinsicness thesis is incompatible with the claim that counterfactual dependence suffices for causation. Taken together they lead to unacceptable results. This can be seen from examples of causation by double prevention, cases in which one event causes another by preventing something that would have prevented the effect from occurring. And indeed the problem of double prevention was one of the reasons that Lewis abandoned the quasi-dependence account. (See [Lewis, 2004a, pp. 83–85].)

(E6) *Double Prevention*: A ball is thrown towards a fragile windowpane. Billy attempts to catch the ball to prevent it from hitting the window, but Suzy trips Billy, preventing him from catching the ball. The windowpane is shattered.

If Suzy hadn't tripped Billy, the windowpane wouldn't have been shattered. So if counterfactual dependence suffices for causation, the tripping counts as one of the causes of the shattering. But note that the causal credentials of the tripping rely partly on factors extrinsic to the actual structure S of events that includes the shattering and all of its causes traced back to the time at which the ball was thrown. For example, the tripping counts as a cause only because of the absence of anything *else* that would have prevented Billy from catching the ball, had Suzy failed to trip him. Suppose Biff had been present and ready to prevent the catch by shoving Billy aside. Then the tripping would *not* have been a cause of the shattering. Yet this new scenario (with Biff present) contains a duplicate of S embedded within it. Hence if we maintain both the Dependence thesis and the Intrinsicness thesis, we will misclassify the new scenario as also being one in which the tripping causes the shattering.

It would appear that either Intrinsicness or Dependence has to go. But Hall suggests a third way out. We may avoid the dilemma if we allow that the concept of cause is not univocal. He argues that there are in fact two concepts of causation [Hall, 2004b]. One of these — *causation as production* — satisfies Intrinsicness and is the concept of causation that the blueprint strategy is intended to elucidate. When we are inclined to judge that the causal relation is transitive, it is causation as production that we have in mind. On the other hand we have *causation as dependence*. When we have the intuition that double prevention is a kind of causation, when we are inclined to think that counterfactual dependence suffices for causation, and when we judge that an absence (the failure of events to occur) can both cause and be caused, it is causation in the sense of mere dependence that we are thinking of.

The blueprint strategy, then, applies only to causation in the sense of production. And it will involve a radical break from the counterfactual tradition. Hall currently suggests that we might use a version of the regularity theory, rather than counterfactual dependence, as a criterion for establishing the causal credentials of a blueprint. See [Hall, 2004a].

Are there really two distinct concepts of cause? One should, in general, be deeply suspicious of this sort of conceptual bifurcation strategy as a way of solving philosophical problems. As Lewis comments:

> ... the many concept hypothesis ... requires distinctions in our thinking that sometimes we do not make, need not make, and are in no position to make. If one event directly causes another, for instance, that is causation in one sense; whereas if one event causes another indirectly, in a case of double prevention (or in some still more indirect case) that is causation in a different sense. But when we neither know nor care whether the causation we have in mind is direct or indirect, what concept of causation are we employing then? [Lewis, 2004b, p. 286].

Lewis goes on to provide an example where we "neither know nor care" whether

we are dealing with production or mere dependence. Air brakes on a train work by double prevention, vacuum brakes by production. Yet the judgment that one may cause the train to stop by pulling the emergency brake cord does not wait upon investigation of the contents of the black box to which the cord is attached. Jonathan Schaffer points out that the trigger mechanism of a gun also typically works by double prevention. No-one, however, is tempted to conclude from this that pulling the trigger of a gun cannot cause it to fire, or only causes it to fire in some secondary sense of "cause". (See [Schaffer, 2000] and [Maudlin, 2004, p. 438].)

ACKNOWLEDGEMENTS

BIBLIOGRAPHY

[Albert and Loewer, 2005] David Albert and Barry Loewer. Counterfactuals and the second law. In *Causation, Physics and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press, Oxford, 2005. To appear.

[Collins *et al.*, 2004] John Collins, Ned Hall, and L.A. Paul, editors. *Causation and Counterfactuals*. MIT Press, Cambridge, MA, 2004.

[Collins, 2000] John Collins. Preemptive prevention. *Journal of Philosophy*, 97:223–234, 2000. Reprinted in [Collins *et al.*, 2004, pp. 107–117].

[Elga, 2000] Adam Elga. Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science*, 68:313–324, 2000. Supplemental volume PSA 200.

[Goodman, 1947] Nelson Goodman. The problem of counterfactual conditionals. *Journal of Philosophy*, 44:113–128, 1947. Reprinted in [Goodman, 1979].

[Goodman, 1979] Nelson Goodman, editor. *Fact, Fiction, and Forecast*. Harvard University Press, Cambridge, MA, 1979.

[Hall, 2004a] Ned Hall. The intrinsic character of causation. In Dean Zimmerman, editor, *Oxford Studies in Metaphysics*, volume 1, pages 255–300. Oxford University Press, Oxford, 2004.

[Hall, 2004b] Ned Hall. Two concepts of causation. 2004. In [Collins *et al.*, 2004, pp. 225–276].

[Hitchcock, 2001] Christopher Hitchcock. The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy*, 98:273–299, 2001.

[Hume, 1902] David Hume. *An Enquiry Concerning Human Understanding*. Clarendon Press, Oxford, 1902. Edited by L.A. Selby-Bigge. Originally published in 1748.

[Lewis, 1973a] David Lewis. Causation. *Journal of Philosophy*, 70:556–567, 1973. Reprinted in [Lewis, 1986a, pp. 159–172].

[Lewis, 1973b] David Lewis. *Counterfactuals*. Harvard University Press, Cambridge, 1973.

[Lewis, 1979] David Lewis. Counterfactual dependence and time's arrow. *Noûs*, 13:455–476, 1979. Reprinted with Postscripts in [Lewis, 1986a, pp. 32–66].

[Lewis, 1986a] David Lewis, editor. *Philosophical Papers*, volume II. Oxford University Press, Oxford, 1986.

[Lewis, 1986b] David Lewis. Postscripts to 'causation'. 1986. In [Lewis, 1986a, pp. 172–213].

[Lewis, 2000] David Lewis. Causation as influence. *Journal of Philosophy*, 97:182–197, 2000. Abridged version.

[Lewis, 2004a] David Lewis. Causation as influence. 2004. In [Collins *et al.*, 2004, pp. 75–106]. Full version.

[Lewis, 2004b] David Lewis. Void and object. 2004. In [Collins *et al.*, 2004, pp. 277–290].

[Mackie, 1965] J.L. Mackie. Causes and conditions. *American Philosophical Quarterly*, 2:245–264, 1965.

[Maudlin, 2004] Tim Maudlin. Causation, counterfactuals, and the third factor. 2004. In [Collins *et al.*, 2004, pp. 419–443].

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

[Schaffer, 2000] Jonathan Schaffer. Causation by disconnection. *Philosophy of Science*, 67:285–300, 2000.
[Schaffer, 2001] Jonathan Schaffer. Causation, influence, and effluence. *Analysis*, 61:11–19, 2001.
[Stalnaker, 1968] Robert Stalnaker. A theory of conditionals. In Nicholas Rescher, editor, *Studies in Logical Theory*, pages 98–112. Blackwell, Oxford, 1968.
[Strevens, 2003] Michael Strevens. Against lewis's new theory of causation: A story with three morals. *Pacific Philosophical Quarterly*, 84:398–412, 2003.
[Yablo, 2000] Stephen Yablo. De facto dependence. *Journal of Philosophy*, 99:130–148, 2000.
[Yablo, 2004] Stephen Yablo. Advertisement for a sketch of an outline of a prototheory of causation. 2004. In [Collins *et al.*, 2004, pp. 119–137].