Causation in a Physical World

Hartry Field

1. Of what use is the concept of causation? Bertrand Russell [1912-13] argued that it is not useful: it is "a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm." His argument for this was that the kind of physical theories that we have come to regard as fundamental leave no place for the notion of causation: not only does the word 'cause' not appear in the advanced sciences, but the laws that these sciences state are incompatible with causation as we normally understand it. But Nancy Cartwright has argued [1979] that abandoning the concept of causation would cripple science; her conclusion was based not on fundamental physics, but on more ordinary science such as the search for the causes of cancer. She argues that Russell was right that the fundamental theories of modern physics say nothing, even implicitly, about causation, and concludes on this basis that such theories are incomplete. It is with this cluster of issues that I will begin my discussion.

Russell's claim that the notion of causation is not needed in fundamental physics has been disputed by Earman [1976], but I think Russell is right and Earman wrong. Earman mentions various causal concepts in physics: determinism, causal signals, and microcausality. But determinism is explainable without the notion of causation, as both Russell in the above article and later Earman himself [1986] have observed. The notion of causal signal is needed in physics only on an operational construal of that; on a less operational view, notions like flow of energy-momentum and various temporal notions such as the light cone structure suffice for the purposes that talk of causal signals have been standardly put. As for microcausality, I'm not sure which of several things Earman had in mind, but I don't see any that support his case.

But of course from the nonappearance of the notion in fundamental physics it doesn't follow that fundamental physics doesn't provide the means to explain it; and it certainly doesn't follow that the whole idea of causation is incompatible with what

fundamental physics tells us. But Russell had two arguments for these stronger conclusions.

One argument (not his main one) concerns directionality. The relation between cause and effect is supposed to have an important temporal asymmetry: causes normally or always precede their effects. This does not appear to be simply a matter of the earlier member of a cause-effect pair being conventionally called the cause; rather, it is connected with other temporal asymmetries that play an important role in our practices. For instance, we tend to *explain* later events in terms of earlier ones but not vice versa; and we think that it makes sense to stop smoking as a teenager so that one will not get cancer later, but that it does not make sense to take a cancer-preventative later in life so that one will not have smoked as a teenager (or to take a cancer-preventative in childhood so that one won't smoke later on). Most people would defend these practices on the grounds that causes explain their effects but not conversely, and that it makes sense to prevent an effect by preventing its cause but not vice versa. The notion of cause is intimately bound up with these asymmetries of explanation and action, as well as with numerous other temporal asymmetries.

But at the level of fundamental physical law, it is hard to see any grounds for the evident directionality of causation. The point is sometimes put a bit contentiously, by claiming that (perhaps with a few minor exceptions) the fundamental physical laws are completely time symmetric. If so, then if one is inclined to found causation on fundamental physical law, it isn't evident just how directionality gets in. But this is an unnecessarily contentious way to put the point: it is not obvious that the claim that the basic laws of physics are time-symmetric is correct; indeed, the notion of the time symmetry of a law itself is not as clear as it sounds.[1]

Russell put the point differently, in a way that doesn't rely on any claim of time symmetry. All the candidates for fundamental laws of physics known at the time he

wrote had the characteristic of being *deterministic in both directions*. That is, from a complete specification of the state of the universe at one time, plus the laws, it follows what the state of the universe is at any other time, *earlier or later*.[2] Russell noted that there seems to be no distinction within fundamental physics between the way in which the past determines the future and the way in which the future determines the past. This seems to be incompatible with the ordinary conception of causation, for part of that conception is that the past determines the future in a more fundamental and important way than any way in which the future might determine the past.

Three points about this argument. The first is that it needn't rely on the claim that the laws of physics are deterministic in both directions, but merely on the idea that they "have essentially the same character in both directions". Giving up determinism wouldn't in itself alter the situation very much, if the laws "had the same indeterministic character in both directions" (to put it vaguely). But *certain kinds of* indeterministic laws might be such as to give rise to a fundamental distinction in temporal direction: for instance, those with well-defined probabilities only in the forward direction. (If one takes quantum mechanical laws to include "the collapse of the wave packet", then the law governing the collapse would appear to be an example; but such "collapse" interpretations of quantum mechanics are highly controversial.)

The second point is that even if one grants that the laws of physics are deterministic in both directions, there is still room in principle for arguing that their character in the forward direction is importantly different from their character in the backwards direction. For instance, Lewis 1979 claims that there are many more determinants of an event at times after the event than there are determinants of it at times before it. I think this claim about an "asymmetry of overdetermination" is wrong, and will discuss it later on.

Even if one can't make a distinction in direction at the level of fundamental law,

one still might make it in physical terms, by bringing in the initial conditions that obtain at our world. Indeed, Lewis's approach was probably intended as an instance of this strategy. But my third point is that this strategy can be made more flexible if one brings in statistical considerations. That is, it might be that there tend to be statistical features of the initial conditions in which we typically apply fundamental physical laws that aren't shared by the final conditions; I will mention possible such features at the end of Section 2 and in Section 4. (The statistical features needn't be supposed to be global features of the universe, they need only be as pervasive as the "directionality" that we observe; so it is no objection to the approach that there might turn out to be a future epoch or a distant region in which these regularities were reversed.) It is such statistical features of the initial conditions that account for the dramatic directionality of the laws of thermodynamics; this gives some initial plausibility to the idea that it might account for the directionality of causation as well.

Price [1992] objects to using statistical considerations to found the directionality of causation, on the grounds that it doesn't give us *enough* asymmetry. In particular, Price argues that the statistical approach doesn't give rise to any important asymmetry in single interactions in conditions where statistical asymmetries are irrelevant–for instance, where only a few particles are involved and they are isolated from their environments (no waves coming in, etc.). I think that Price is partially correct: in such situations, there is no asymmetry that is intrinsic to the interaction itself. Still, if the *earlier than* relation is associated with certain statistical regularities (even ones local to our epoch and our region of the universe), then in appealing to this relation in situations with no intrinsic statistical asymmetry one is still invoking an extrinsic statistical asymmetry; and I think that a defender of the statistical approach can say that this is the only temporal asymmetry there is reason to believe in. This would mean that within certain systems, an explanation of the past state by means of the future state is intrinsically on par with an explanation of its future state by its past state: preference for the latter over the former could be justified

only by appeal to other systems in which the statistical regularities matter. So the statistical approach may not give enough asymmetry *to validate our ordinary preconceptions*; but perhaps the problem is with the preconceptions, not with the statistical approach.[3]

If something like this is right, then Russell's first argument is problematic: although it is true that the notion of 'cause' is not needed in fundamental physics, even statistical physics, still directionality considerations don't preclude this notion from being consistently added to fundamental physics; and indeed, it may even be the case that the notion can be explained within statistical physics. Such an explanation would not capture our full intuitive preconceptions about the directionality of causation, but it could capture a good bit of them.

But Russell had another argument, on which he put more weight. He claimed that our causal way of thinking relies on the assumption that there are laws that tell us that when a finite number of quite localized things hold at one time, some other particular thing must happen a short time later. (When someone strikes a non-defective match and holds it to a flammable substance and there's oxygen present and a few other things hold, a fire must result.) Russell points out that no proposed "law" of this sort has a chance of being correct,[4] and that physics has progressed by replacing such alleged laws by differential equations. In some ways differential equations have a very different character: for instance, instead of directly connecting things at two different times (which leaves lots of opportunities for outside influences to make things go wrong), a differential equation involves a single time only: it determines the rate at which a quantity changes at a given time t from the value of it and other quantities at that very time; by giving the rate of change, it indirectly gives you the values at other times, though only when a very detailed description of the values at t are plugged in. In fact, even when one assumes that "causal influence" can't exceed the speed of light, still one will need a description of an entire cross-section of the past light cone of an event to determine the event. Somewhat

more precisely, *information about what happens at an earlier time can't suffice to determine the event unless it includes information about each point at that time that is within the past light cone*; only when there is information about each such point can the possibility of intervention from afar (e.g. by extremely powerful pulses of energy) be excluded. This seems to mean that (assuming determinism) facts about each part of the past light cone of an event are among the causes of the event.[5] (Of course, most such facts won't be salient enough to be worth mentioning in typical contexts where we are asked to cite causes, but like the presence of oxygen when (or a moment before) a fire starts, they are causes nonetheless.) Russell did not consider the possibility of indeterministic laws, but the point would be little changed if he had: the general point is that no reasonable laws of physics, whether deterministic or indeterministic, will make the probability of what happens at a time depend on only finitely many localized antecedent states, one will need an entire cross-section of the light-cone to make the determination. Indeed, given quantum nonlocality, one will need even more.

Perhaps all this shows is that an event has a lot more causes than we may naively assume; what's the big deal in that? But there would be a big deal if we had to conclude that if c_1 and c_2 are both in the past light cone of e then there is no way of regarding one of them as any more a cause of e than the other: then Sam's praying that the fire would go out would be no less a cause than Sara's aiming the water-hose at it, and the notion of causation would lose its whole point. *One* way to read Russell (possibly not the most interesting way) is as implicitly arguing that the form of our physical laws makes this conclusion inevitable; such an argument is explicitly discussed by Latham 1987, and I think it raises a serious problem for those, like Davidson 1967, who restrict causes to fairly concrete events.

More explicitly: what I take to be the clear truth behind the (Russell?)-Latham argument is that since there is always a possibility of interventions from afar, the non-occurrence of those interventions must be included among the causes of an event. Instead

of sitting there idly praying, Sam might have taken effective means to keep the fire going, say by shooting a hole in Sara's water-hose before it could put out the fire. His *not* shooting the hose should be included among the causes of the fire going out. Of course, this is the kind of cause that we wouldn't mention (unless there was some special reason to think he *might* shoot the hose); like the presence of oxygen in the lighting of a match, it is an extraordinarily non-salient cause, but it is a cause nonetheless, and virtually every serious account of causation will treat it as such. This is not in itself a problem. But it would become a problem if we thought that causes have to be events and that when Sam was sitting there praying instead of shooting the hose, there was only one event (a praying-and-not-shooting-the-hose event): for then we would have to conclude that his praying was a cause of the fire going out. And by extension of the reasoning, we would have to conclude that everything about the past light cone of the fire's going out was a cause of it. To avoid this, we better avoid the Davidsonian view that only quite concrete events can serve as causes: we should instead say either that facts as well as events can serve as causes (Bennett 1988); or that the events that serve as causes can be highly unspecific, including "omissions" like Sam's not shooting the hose (Lewis 1986c, 1986a); or some such thing.[6]

I don't think Latham's argument is all that Russell was worried about when he stressed the difference between differential equations and simple pre-scientific laws about what happens when you strike a match. Whether his other worries are more troublesome I'm not sure. (I'll mention one or two of them later.) At any rate, Russell's conclusion was that the notion of causation is hard to make sense of in physical terms, and from this he drew the conclusion that it is a notion that we ought to abandon. Abandoning it would do no harm, he thought, because physics doesn't need it.

2. But as Cartwright points out, the cost of abandoning the notion of causation is intolerably high: for that notion is intimately connected with the distinction between effective and ineffective strategies. We all think that for the goal of avoiding lung cancer,

it's beneficial to stop smoking. Intuitively the reason is that smoking is a cause of lung cancer. The reason is *not* simply that there is a high statistical correlation between smoking and lung cancer. For correlation is symmetric: if there is a high statistical correlation between smoking and lung cancer, there is a high statistical correlation between lung cancer and smoking. But for the goal of stopping smoking, it is not in the least beneficial to take a cancer-preventing drug, because cancer isn't a cause of smoking. Similarly, there is a high statistical correlation between lung cancer and the foul breath that cigarettes produce, due to the fact that both are caused by smoking. But for avoiding cancer, breath mints do no good; nor would a cancer-preventing drug be likely to be of use in avoiding bad breath.

The most dramatic illustrations of the point also illustrate something called "Simpson's paradox", a surprising fact about statistics known since about the turn of the century, but which before Cartwright's article was not known as widely among philosophers as it should have been. The illustration that follows is not hers, but derives from a discussion of Ronald Fisher's of whether we have evidence that smoking causes cancer. I've added a small twist.

Imagine that we have performed a statistical study in which many people are randomly chosen and studied over the course of a lifetime; they are categorized in terms of whether they smoked heavily in their early years and whether they got lung cancer later in life. Imagine that the breakdown is as follows:

	Smokers	**Non-smokers**
Cancer	49,501	9,980
No Cancer	50,499	890,020

Total	100,000	900,000

It looks like if you smoke, your chances of cancer are 49,501/100,000 = 0.49501, whereas if you don't your chances are only 9,980/900,000 = 0.01109. "Obviously you're much better off not smoking." (By a factor of about 45 to 1.)

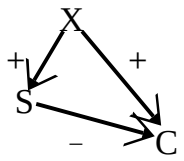
But to continue the fantasy, suppose we gather further information about these very same people, namely, information as to which ones possess a certain gene, Gene X. (This new information in no way alters the statistics above.) Let's pretend that this new information allows us to break down the original table as follows:

	Smokers w X	**Smokers w/o X**	**Non-smokers w X**	**Non-smokers w/o X**
Cancer	49,500	1	990	8,990
No Cancer	49,500	999	10	890,010
Total	99,000	1,000	1,000	899,000

Suppose you have Gene X. Then if you smoke, your chances of cancer are 49,500/99,000 = 0.5, whereas if you don't your chances are 990/1000 = 0.99. "Obviously if you have Gene X, you're much better off smoking." (By a factor of almost 2 to 1.) But suppose you don't have Gene X. Then if you smoke, your chances of cancer are 1/1000 = 0.001, whereas if you don't your chances are 8,990/899,000 = 0.01. "Obviously if you don't have Gene X, you're also much better off smoking." (By a factor of 10 to 1.) The information about the gene in no way alters the overwhelmingly high statistical correlation between smoking and cancer, but seems to dramatically alter its significance. For the natural conclusion from the second data is this: *Smoking is not a cause of cancer, and in fact tends strongly to prevent cancer; though there is a strong positive statistical correlation*

between smoking and cancer, that is because they have a common genetic cause.[7] So from a health-conscious point of view, one ought to endure the disgusting habit of smoking because of its health benefits.[8] (Perhaps it even is a health benefit to all those other people who are forced to breathe your second hand smoke.)

The example illustrates three points. First, it emphasizes what should have been a familiar point anyway, that correlation doesn't imply causation. (The earlier examples of lung cancer not causing smoking and of neither lung cancer nor nicotine breath causing the other already made this clear.) Second, it illustrates the initially surprising mathematical fact that *a variable S can be positively correlated with a variable C overall and yet be negatively correlated with C both conditional on a third variable X and also conditional on* $\neg X$. The example not only illustrates this, but shows how it can happen: it can happen if the causal situation is as pictured in the following diagram (where the labelled arrows indicate positive or negative influence),



and if the positive correlation between S and C induced by their common cause X is sufficient to outweigh the negative correlation between them that results from the preventative effect of S on C. The third point that is illustrated is that it is the causal conclusions and not the correlations that we need to know in order to best achieve our ends. (Cartwright concedes that we might be able to make do with probabilistic notions instead of causal notions, but only on a conception of probability that is "causally loaded", i.e. that amounts to something like "the probability obtained by holding all causal factors fixed".)

I think this makes a compelling case against Russell's view that we should do without causal notions. But Cartwright herself draws a much stronger conclusion, a kind of *causal hyper-realism*, according to which there are causal facts *that outrun the totality of "noncausal facts"* (i.e. the facts that could be expressible in some language without using causal terminology). Indeed, her claim isn't simply that there is no reasonable way to explicitly define causation in noncausal terms; it seems to be that causal claims don't even supervene on the noncausal facts. Among the "noncausal facts" she includes the basic laws of physics–e.g. Newton's law that an object accelerates in direct proportion to the force impressed on it and in inverse proportion to its mass. She holds that the causal fact that a force on an object *makes* the object go faster is not reducible to Newton's law, nor to other noncausal facts either, such as the equations of energy flow from the sources of fields to the fields themselves to the accelerating objects. (Such equations are just further parts of fundamental physics, which she regards as "laws of association" rather than as causal.) Rather, the claim that a force on an object makes the object go faster states a further truth about the world that physics leaves out. Evidently there is some sort of causal fluid that is not taken account of in the equations of physics; just how it is that

we are supposed to have access to its properties I am not sure.[9]

But despite the implausibility of the hyper-realist picture, we have a problem to solve: the problem of reconciling Cartwright's points about the need of causation in a theory of effective strategy with Russell's points about the limited role of causation in physics. This is probably the central problem in the metaphysics of causation.

One thing that needs to be noted about this problem is that the examples given of the need of the notion of causation (for instance, 'Teenage smoking tends to cause lung cancer') have concerned *general* causal claims among variables that are *fairly inexact* in the sense that they can be instantiated in many different ways. First let's discuss the generality. The precise connection between such general causal claims and specific causal claims like 'Joe's teenage smoking was a cause of his lung cancer' is complex and controversial: some have thought that singular causal claims should be explained in terms of general causal claims, some have thought the order of explanation should be reversed, some have thought that both should be explained in terms of some third thing (for instance, objective probability), and some have thought that we simply have two different kinds of causal claims that are only loosely connected with each other.

The view that takes general causal claims as primary is totally implausible if general causal claims include claims like 'Turning the left knob on radios clockwise causes the volume to increase'. That may be a true generalization, or may have been one at a time when radio cases had a certain design, but as a generalization over systems of different sorts it has little interest or robustness. General claims *about a given system*, such as 'Turning the left-most knob of this radio causes the volume to increase', might with more plausibility be regarded as prior to singular causal claims.

But "singularist" views that take singular causal claims as primary and explain general claims in terms of singular seem at least equally plausible.[10] On such a view, 'Smoking is a cause of cancer' might mean something like "The probability of a person's

smoking being a significant cause of his getting cancer, given that he smokes and given an appropriate context, is not insubstantial."[11] Should the fact that the motivation we have given for the need of causation involves general causal claims undermine this? I don't think so: for in fact what is important to know for deciding how to act isn't whether refraining from smoking is *generally* an effective strategy for avoiding cancer, but whether it would be an effective strategy *for me* to adopt. And so what I need to know is the chance that smoking will cause cancer in my own case. The focus on generalizations was really just a matter of convenience, justified only in circumstances where the agent can regard himself as a typical member of the population.[12]

But though it is probably not significant that our examples of the need of causation have concerned *general* causal claims, I think it is significant that they have concerned claims that involve *fairly inexact variables*. 'Inexact' here doesn't mean 'vague': rather, a variable is inexact if the claim that it assumes a given value on an occasion can be realized in many different ways that on a deeper level of analysis are importantly different. If I am deciding whether to smoke, then even if I have detailed information about the other factors that are relevant to whether I will get lung cancer, I certainly can be nowhere near having *enough* information: the outcome is bound to depend on fine details about the state of my body now and of the rest of the universe that I will interact with, and on the details of how I might carry out my decision to smoke or to refrain from smoking. (It will also depend on the outcomes of irreducibly chance processes, if the universe is indeterministic; but it is the kind of statistical probability that exists over and above any ultimate chanciness that there may be that is important to my current point.) This means that the predictions of interest to us could not be made on the basis of the underlying physics without the use of substantial statistical assumptions, of the general sort that are also required for thermodynamics. The notion of causation, like the notions of temperature and entropy, derives its value from contexts where statistical regularities not necessitated by the underlying physical laws are important. As noted before, that does

not necessarily mean that the notion of causation can't be applied in contexts where such statistical regularities are absent; but it does make the point of causal talk in such contexts depend in a surprising way on factors extrinsic to the contexts.

I will conclude this section by mentioning a recent account of the empirical confirmation of causal generalizations by correlational data (Spirtes, Glymour and Scheines 1993; see also Pearl 2000). The SGS account has a considerable bearing on the topics I have discussed. It makes clear that causal graphs like the one displayed several pages back play an important role in our thinking about causation and about how we expect the correlations to alter when we make decisions (or when the system is disturbed from the outside); and it makes explicit the assumptions we employ about how the causal structure of the graph constrains the assignment of objective probabilities to combinations of values of the variables. In doing so, it makes clear precisely *how* the notion of causation is directional. If Z is a common *cause* of X and Y, then that will tend to induce an *unconditional* probabilistic dependence between X and Y; though as long as there are no intermediate common causes, holding fixed the value of Z cancels the probabilistic dependence of X and Y.[13] If on the other hand Z is a common *effect* of X and Y, it is pretty much the other way around: the common effect *doesn't* induce an *unconditional* dependence between X and Y, but holding the common effect fixed does induce a probabilistic dependence.[14] (The account of how probabilities are to be modified when the system is disturbed by a decision or from a "natural" occurrence from the outside also involves directional elements, but this asymmetry seems to derive from the basic asymmetry in the way that causal graphs constrain probability. See the "Manipulation Theorem" in Spirtes et al, section 3.7.2.)

This account of the asymmetry in causal graphs gives a way to make fairly precise the temporal asymmetry that underlies the concept of causation: the asymmetry consists in the pair of facts

(M) that the variables we find salient tend to be probabilistically related in such a way that you can draw causal graphs among them in accordance with the Spirtes, Glymour and Scheines conditions (the ones roughly sketched in the preceding paragraph);

and

(T) (Around here anyway) what are causes on the causal graph criterion tend to precede their effects.

(This is similar to, though more comprehensive than, the "fork asymmetry" discussed by many writers, e.g. Horwich 1987.) The salience condition needs emphasis: if the universe is two-way deterministic as in classical physics, one can find very unnatural variables for which the temporal orientation in (T) is reversed: see Arntzenius 1993 (secs. 5 and 6). And with "exact" variables in the sense explained above, the asymmetry completely disappears in classical physics. Quantum mechanics gives rise to more dramatic failures of (M) and (T) together: the much-discussed nonlocal correlations between measurements of distant particles[15] cannot be given explanations that accord with (M) except with highly unnatural causal graphs (for instance, ones where the outcome of the measurement influences either the prior state of the particles or the prior choice of settings of the measurement instruments). But despite all these limitations on their scope, (M) and (T) together describe an overwhelmingly pervasive asymmetric regularity on the macroscopic scale. Interesting questions arise about how this regularity is to be explained, but I will not pursue them.

Even if there are extraordinarily non-salient variables for which what would count as a "cause" by the SGS constraints comes later than what counts as an "effect", it doesn't follow that we should take the SGS approach as dictating that some effects precede their causes. A better conclusion is that causes must always be temporally prior to their effects, but that the non-arbitrariness of this is revealed by the fact that it accords with

what is required for causal graphs among salient variables that obey the SGS constraints. (Horwich 1987 makes a similar suggestion.)

The SGS theory may confirm one of Russell's suspicions: the causal graphs that its authors employ involve only finitely many variables, and this fact[16] plays a key role in how they develop the theory. If the theory *can't* be developed independently of this assumption, Russell would appear to be right in holding that the methodology of testing *general* causal claims essentially requires a radical idealization of the underlying physics. And perhaps this conclusion could be transferred to singular claims too, if causal graphs play a substantial role in the theory of them (as I think is likely). However, I don't think it at all obvious that the causal graph approach can't be generalized. Intuitively, it seems (barring quantum nonlocality and the like) that one should be able to think of the physical universe as a causal system with a node for each space-time point, with the value of the node expressing the totality of the values of physical quantities at that point; the light-cone structure gives the dependence relations. The kind of simple causal systems we employ in practice seem as if they ought to have such a "non-discrete causal system" as a limiting case. But of course the details of this vague suggestion would need to be worked out, and I wouldn't be surprised if some of our causal intuitions (e.g. about preemption, soon to be discussed) would fail to be validated in the limiting case.

3. I have emphasized the statistical underpinnings of the notion of causation, at least with regard to directionality; but that does not necessarily mean that an account of causation applicable to individual processes must make explicit reference to statistical facts. How else might an account of singular causation proceed?

An idea mentioned in passing by Hume (1748), and taken up by Lewis (1973), is that causation involves counterfactual dependence. Hume and Lewis use 'would' counterfactuals: the initial idea (before the bells and whistles are added) is that for John's smoking to have caused his cancer, it must be the case that if he hadn't smoked he

wouldn't have gotten cancer.[17] An alternative (McDermott 1995a) uses 'might' counterfactuals: the initial idea is that for John's smoking to have caused his cancer, it must be the case that if John hadn't smoked he might not have gotten cancer; that is, it must not be the case that if he hadn't smoked he would have gotten cancer.[18] There are also intermediate alternatives, e.g. 'would probably' counterfactuals and more complicated counterfactuals that involve comparisons of probability. The differences between these alternatives are not insignificant (and they arise under determinism as well as under indeterminism),[19] but will not be discussed here.

The counterfactual framework is broad enough to encompass many accounts of causation not often thought of as counterfactual accounts. For instance, Mackie (1965) has offered an account of causation under determinism in terms of which for C to cause E, there must be a minimal sufficient condition for E that includes C (minimally sufficient given the basic physical laws). Given determinism, the existence of a sufficient condition for E that includes C is trivial; what's crucial is the minimality, and what it says is that excluding C from the condition, the laws no longer guarantee E.[20] A natural way to put that is: if C hadn't occurred, E might not have either. On a suitable account of counterfactuals (basically, the Goodman account in terms of laws), Mackie's account is just an account in terms of 'might' counterfactuals.

Counterfactuals are of course notoriously context-dependent: much more so even than causal claims. It is perfectly within the bounds of ordinary counterfactual talk to say that if the barometer needle hadn't dropped, there wouldn't have been a storm a short time later; but no one wants to say that the dropping of the barometer needle was a cause of the storm. Similarly, it is perfectly acceptable to say that if Jane's parents had both had blue eyes, Jane would have had blue eyes too; but most people who say that know that this is not a cause-effect relation but is due to a common cause, the parental genotypes. Lewis recognizes that such counterfactuals must be discounted if causation is to be based on counterfactuals. Lewis says that the barometer counterfactual and the blue-eyes

counterfactual are acceptable only in special contexts, contexts where we allow "backtracking arguments". If a backtracking argument is one where we reason from effects to causes, it is plausible to say that such counterfactuals do involve backtracking arguments: a scientifically knowledgeable person asked to defend the barometer counterfactual would do so by saying that if the barometer needle fell that would most likely be due to a region of low pressure that would be likely to cause a storm. It is also plausible that there are contexts in which we do not accept counterfactuals supported only by backtracking arguments. I don't know if Lewis is right that the latter contexts are "the normal ones", but doubt that it much matters: it would be no serious threat to a counterfactual account of causation if it required a somewhat specialized kind of counterfactual. What does matter, though, is whether the distinction between the two kinds of contexts can be made without appeal to the notion of causation. If it can't, then there seems to be a circularity in a counterfactual theory that depends on the restriction to non-backtracking counterfactuals (counterfactuals that can be defended without appeal to backtracking arguments).

One possible way to avoid the use of the notion of causation in distinguishing backtracking counterfactuals from others is to use temporal order instead. Of course, in the above examples it can't be used directly: the barometer counterfactual and the blue-eyes counterfactual involve the same standard time order present in the case of non-backtracking counterfactuals (the time of the consequent is later than that of the antecedent).[21] But perhaps we could argue that these counterfactuals are only supportable by means of other counterfactuals (a barometer to low pressure counterfactual or a parental phenotype to parental genotype counterfactual) that involve reverse time order. We would have to argue this *without using causal notions* if this was to help the counterfactual theorist of causation; but maybe this could be done.

This approach is reasonably attractive, but it is not one that Lewis can use. For one of the main advantages that Lewis claims for his counterfactual approach to causation

is that it *explains* the directionality of causation: more specifically, it explains why causes nearly always precede their effects (at least in our part of the universe in the current epoch), and does so not merely as a result of stipulation that causes are always prior but in a fashion that illuminates the genuine directional asymmetry noted earlier. But he could not claim this advantage for the counterfactual account if his account of causation were to be based on a restriction to those counterfactuals in which the time of the antecedent precedes the time of the consequent: then causes would precede effects simply by fiat.

It is sometimes suggested that the "peculiarity" of backtracking conditionals arises from their extreme indeterminacy or uncertainty. If this were so, perhaps excluding the extremely indeterminate or uncertain counterfactuals would suffice to exclude backtrackers without relying either on explicit fiat or on the notion of cause. But the claim that the peculiarity of backtrackers arises from their extreme indeterminacy or uncertainty appears to be false. Consider the following pair:

(1) If Oswald hadn't killed Kennedy in 1963, Kennedy would have won the 1964 election.

(2) If Kennedy had won the 1964 election, Oswald wouldn't have killed him in 1963.

Provided that we don't simply exclude backtracking conditionals, (2) seems more certain than (1) (and less likely to be indeterminate in truth value than (1)): with (1), there's always a chance that the affair with Marilyn Monroe would have become public and that this would have outraged the American public so much that they preferred Goldwater (or whoever the Republican candidate might have been); a comparable story for (2) would have to be wilder (the government keeping his assassination secret; or the public being so outraged by the choice of Johnson and Goldwater that they preferred to write in a dead man; or whatever). I agree that many kinds of counterfactuals seem more indeterminate in the backwards direction than in the forward, but this and many other examples make it highly doubtful that our tendency to exclude backtrackers can be wholly explained as due

to that fact.

Lewis makes a different attempt to found the distinction between backtracking and non-backtracking counterfactuals independently of both causation and time, but it too is a failure, as I will argue in the next section. Still, I think it may be possible to found the distinction using less than the full notion of singular causation, and in a way that allows for a serious explanation of directionality.[22] Let us put aside any doubts we may have that this can be done, and return to how non-backtracking counterfactuals can be used in an account of causation.

Suppose C and E are true event-statements about disjoint regions. The simplest 'would'-counterfactual approach to singular causation would be to take causation to be counterfactual dependence (in a non-backtracking sense): (The fact that) C is a cause of (the fact that) E iff if C weren't the case, E wouldn't be either; or in symbols, $\neg C \mathrel{\Box\!\!\rightarrow} \neg E$.[23] This was essentially Hume's proposal. (Not his main view, of course, but a proposal he mentioned in passing.) Lewis (1973) weakened the account slightly: C is a cause of E iff there is some chain of true event-statements $A_0, A_1, ..., A_n$, with A_0=C and A_n=E, such that for each $i<n$, $\neg A_i \mathrel{\Box\!\!\rightarrow} \neg A_{i+1}$. This modification of Hume guarantees that singular causation is transitive.

There is however reason to doubt that singular causation should be thought transitive. Consider a famous example from Cartwright 1979 (one that Cartwright uses for a different purpose). Suppose that Nancy sprays a fairly effective weed-killer on a weed in her garden; this triggers its "immune system" to counter it, and as a result the plant survives to photosynthesize for years to come. The spraying of the weed-killer was a cause of the "immune reaction"; the immune reaction was a cause of the survival, or of the future photosynthesis; so transitivity dictates that the spraying of the weed-killer should be a cause of the survival, or of the future photosynthesis. Pre-theoretically, this seems dubious. Further examples casting doubt on transitivity can be found in

McDermott 1995b. The examples are probably not decisive, but they certainly have some force, and raise the question of whether there are good reasons to accept transitivity.

Lewis's reason for accepting transitivity (not only in 1973 but in two substantial revisions of his view, Lewis 1986 and Lewis 2000) was to deal with a certain sort of preemption. Suppose Joe throws a rock at a window, breaking it. Pete, closer to the window, was poised to throw an identical rock along the final segment of the same path that Joe's rock actually took, in a way that would reach the window at the same time with the same velocity; he refrained from doing so because he saw Joe's rock coming. We want Joe's throwing the rock to count as a cause of the window breaking; but if he hadn't thrown the rock, the window would have broken anyway (and in just the way it actually did). We have causation without counterfactual dependence.[24] But if we accept the transitivity of causation, we will get the desired causal claim as long as there are intermediate events that counterfactually depend on Joe's throw and on which the window's breaking counterfactually depends. And there are such, though we must be a bit careful in how to choose them.[25]

I'm skeptical, though, that the appeal to such intermediate events reflects our intuitive rationale for judging that Joe's throwing the rock was a cause of the window's shattering. The intuitive rationale, I think, involves *conditional* counterfactual dependence (rather than chains of unconditional counterfactual dependence): the intuitive rationale is that *holding fixed the fact that Pete didn't throw*, the window's shattering does depend counterfactually on Joe's throw.[26] If we can explain causation in terms of *conditional* counterfactual dependence, we don't need to transitivize. It seems to me that there is something quite odd about the use of transitivity to handle such examples of preemption, for such preemption examples seem intimately related to *failures* of transitivity: if Joe hadn't thrown, that would have caused Pete to throw, which would have caused the window to break; so transitivity tells us that Joe's not throwing would have caused the window to break! Moreover, there are many cases of preemption that

invoking transitivity clearly won't solve (the various sorts of "late preemption" discussed in Lewis 1986a), but many and perhaps all of them seem naturally handleable in terms of conditional dependence.[27] The suggestion that we use conditional dependence rather than transitivity to handle preemption is plausibly developed in Hitchcock (forthcoming). Hitchcock notes a different reason for finding the transitivity approach implausible: in cases (like the window case) where transitivity yields intuitively desirable results it is only by virtue of extremely carefully chosen intermediate variables (see note 25), so that the application of transitivity is unobvious; whereas in cases (like the Cartwright case) where it yields intuitively undesirable results, the fact that transitivity yields those results is plain to see. If transitivity is what is responsible for our intuitive judgements of causation, we ought to find causation obvious in the Cartwright case and much less obvious in the window case.

I don't mean to suggest that all problems about preemption are solved merely by adopting the conditional dependence approach. Collins 2000 presents an interesting group of puzzle cases that the approach doesn't seem to handle as it stands. Moreover, Cian Dorr has pointed out to me that with a clever choice of variables to hold fixed, many of the dubious cases of causation that are clearly licensed by transitivity can be argued to be licensed (though at least less obviously) by conditional dependence as well. The Hitchcock paper discusses this as well, and proposes a way to deal with it. A fuller discussion of the adequacy of his resolution or other possible resolutions would probably connect up with an issue from Russell with which I began: is causation a notion that has application only within an idealized model of the world, in which we simplify the causal history of an event by focussing on a few "causal pathways" and ignore interactions among them and interventions from outside, or should we rather give an account that is sensitive to the fact that some features of every point in the past lightcone of an event is causally relevant to that event? In addition to this issue, issues about independent complications in the account of causation, e.g. to handle symmetric overdetermination

and causation under indeterminism, are sure to enter in. Such further discussion is far beyond the present scope.

4. I now return to the issue of the directionality of causation. Lewis (1979 and 1986b) claims that a main advantage of (his version of) the counterfactual account of causation over what he calls "regularity analyses" is that the former can straightforwardly account for causal directionality. If we presuppose determinism and confine attention to causal claims where the cause statement C involves a certain instant of time t_1 and the effect statement E involves a certain distinct instant of time t_2, a simple "regularity analysis" might say that C is a cause of E if and only if there is a minimal condition C* involving t_1 that suffices for E given the basic physical laws, and C* entails C. But unless we are willing to add the further requirement that t_1 precedes t_2, no account much like this can work in a 2-way deterministic universe, since the right hand side will be equally true when E is a cause of C.[28] And Lewis thinks it is a major defect in a theory of causation that it builds in the condition that the time of the cause precede that of the effect: that causes precede effects is something we ought to *explain.* A main advantage that he claims for his counterfactual theory is that it explains it.

Since Lewis explains causation in terms of counterfactual dependence, the issue for Lewis is how to explain the asymmetry of counterfactual dependence. If causation is explained in terms of counterfactual dependence, counterfactual dependence must be explained without use of the (full) notion of cause, and the objection to building in a temporal precedence requirement into causation by fiat means that we can't simply rule out counterfactuals where the time of the antecedent is later than the time of the consequent by fiat. We're back to the issue of how to rule out backtrackers. (We have already observed that Lewis needs to count "common cause" counterfactuals as backtrackers, even when the time of the antecedent precedes that of the consequent, so temporal fiat wouldn't directly suffice anyway, in the case of counterfactuals. But as I mentioned, such a fiat might be thought to rule out common cause backtrackers

indirectly.)

Lewis offers a very ingenious attempt to rule out backtrackers without temporal fiat, by an account of similarity among possible worlds. According to his basic account of counterfactuals, ¬C □→ ¬E is true iff there is a ¬C-world w (a possible world w where ¬C holds) such that every possible ¬C-world that is at least as similar to the actual world as is w is a ¬E-world. He then proposes an account of similarity among worlds. If determinism holds in the actual world (as I will continue to assume for simplicity), then this account of similarity, taken together with familiar facts about the actual world, is supposed to have a dramatic consequence: that when C is a true event-statement such that we might seriously entertain ¬C as the antecedent of a counterfactual, there will be ¬C-worlds not too far from actuality, *and all of them will be like the actual world up until a time very shortly before the time of C*. At that point in any such world, a small miracle occurs, i.e. a small violation of the laws of the actual world, so as to allow ¬C to hold; but then immediately afterwards, the laws of the actual world remain unviolated. This, I repeat, is supposed to follow (given contingent facts about the actual world) from an account of similarity *that doesn't rely on causal notions and doesn't build in a specific direction of time*. The similarity ordering of worlds that satisfies this account is the one we assume when we exclude back-trackers; for back-trackers there is a different similarity ordering, which we need not consider. So non-back-tracking counterfactuals, Lewis claims, are simply counterfactuals that are to be evaluated by the standard similarity ordering.

Before discussing the main problem with this account, I'd like to note an immediate oddity about it. The oddity is that by avoiding temporal and causal notions in explaining backtracking, Lewis hasn't really *eliminated* backtracking in the normal sense (the sense where a true counterfactual of the form ¬C □→ ¬E where the time of E precedes the time of C is always a backtracker), but merely *limited* it. That is: assuming determinism, Lewis's account has it that it would have required a miracle for Princeton

not to have made the job-offer to him that they in fact made, at some time t in 1969; and that the miracle would have had to occur at some point prior to t. (I'm assuming that we restrict our attention to worlds whose remote past is like that of the actual world, as Lewis recommends.) It may not be determinate precisely which miracle would have occurred before t, but it is determinate that the world before t would have had to be a bit different somehow. So Lewis's account would seem to lead to the result that Princeton's offering him the job was a cause of some (highly conjunctive) effect just prior to it. (Lewis responds to this in Lewis 1981; for a counter-response, see Vihvelin 1991.) Indeed, for other counterfactuals, a price of keeping the miracle small is that the backtracking isn't limited to a very small time before (Bennett 1984): there are nonvacuous counterfactuals about what would have happened if Goldwater had won the 1964 election, but it is hard to see how a small miracle *a second or so before the time of the election* could have altered the outcome; there is a substantial period prior to the election that would have had to have been different for Goldwater to have won, and given Lewis's approach to backtracking some facts about this whole period would seem to come out as effects of Goldwater's losing. Anomalies like this seem to show that the prospects of explaining the intuitive distinction between back-trackers and non-back-trackers without appeal to either causation or a temporal direction are dim.

But a more fundamental worry about Lewis's account is that nothing in his account of similarity among worlds seems as if it can possibly explain why the "small miracle" in the worlds most similar to the actual world must happen just prior to the time of the antecedent rather than just afterwards. (If it happened just afterwards, the possible worlds in question would be just like the actual world at later times, and differ drastically at times before, so that effects would mostly precede their causes rather than succeeding them.)

Lewis *appears* to address this point, when he argues that when you take a possible world like ours in its initial stages, and introduce a small miracle in it at a time t_1, then if the world operates by the normal laws until a later time t_2, you will need a much bigger

miracle at t_2 to bring the world back to coincidence with the actual world (because of the way that small differences get amplified over time). Lewis concludes that this illustrates a temporal asymmetry: because of something about what our world is like, "convergence" miracles that bring worlds hitherto unlike ours into line with ours must be bigger than "divergence miracles" that make worlds initially like ours start to differ. But this conclusion is too rash: the fact that you need a big miracle at t_2 to make a world that has diverged from the actual world at t_1 *reconverge* is really part of a temporally symmetric fact. The temporally symmetric fact that emerges from Lewis's discussion is that in a two-miracle world that is just like the actual world in both initial and final stages (and where there is a more than minuscule time between the two miracles), the miracles can't *both* be small; but the initial (divergence) miracle could be the big one.

The real issue for Lewis's account isn't the respective sizes of divergence miracles and *re*convergence miracles; it's the respective sizes of (i) divergence miracles at a time $t_1 - \epsilon$ just prior to the actual time t_1 of C, in ¬C-worlds that are just like the actual world up until $t_1 - \epsilon$, but which may differ drastically after t; and (ii) convergence miracles at a time $t_1 + \epsilon$ just after the actual time of C, in ¬C-worlds that are just like the actual world after $t_1 + \epsilon$, but which may differ drastically before t. And (unless C is itself a temporally loaded claim) it is clear that the size of the miracle can be equally small in the two cases. Suppose for instance that C is a true claim about the position of one or more particles at t_1. One way to imagine a world that obeys the same laws as ours except for a minor miracle near t_1 is to suppose that the world is exactly like ours up until $t_1 - \epsilon$, but with shifted positions for these particles at t_1; the operation of the normal laws after t_1 will ensure that the world will become extremely different from ours at much later times. But just as easily, we can imagine that the world is exactly like ours after $t_1 + \epsilon$, but with shifted positions for these particles at t_1; the operation of the normal laws before t_1 will ensure that the world was extremely different from ours at much earlier times. There is simply no asymmetry here of the sort Lewis claims.[29]

Lewis's discussion of the asymmetry of traces may appear to provide an argument in the other direction. In our world, a stone dropping in a pond leaves many traces, in ripples proceeding outward and light waves being sent off into space and in the memories of those watching. What small miracle in a world where the stone didn't drop could have produced all these effects? It is important to realize that since the miracle-world need not (and can not) be a *re*convergence world, then in the period just prior to the miracle as well as in the period afterward, it can have misleading apparent traces of the stone having dropped: it needn't be the miracle that produces the misleading apparent traces. This removes a major intuitive obstacle to the miracle being small.

Even so, it may seem hard to conceive of what such a miracle-world would look like. But Elga (forthcoming) points out that if we assume the laws of physics to be time-reversal invariant, there is an easy way to conceive this. Take the time-reverse of our world. (Roughly, the world that obeys the same physical laws as ours, and is now just like ours is now except with all particles having the opposite direction of travel.)[30] This world "looks exactly like ours run backwards"; the concentric waves of water and light rush inward to converge on the stone as it reaches the surface of the water from below, propelling it upward. This seems an amazing and improbable coincidence; but the fact that it is compatible with basic physical laws, though statistically improbable, is uncontroversial. (The fact that cases like this never or virtually never happen, though their time-reverses are common, is a fact that Lewis calls "Popper's asymmetry"; he is quite explicit that it is a matter of statistics only, not fundamental law.) It seems intuitively clear that a small miracle in the time reversed world just before the stone was propelled upward (but after the incoming waves were noticeable) could have destroyed the delicate balance in initial conditions required for the stone to leave the surface. But the time-reverse of this is a world where the rock doesn't fall to the surface, but leaves the apparent traces of having done so.

Of course, it is clear that one-miracle worlds that are like ours in their final stages

but not in their earlier ones will be extremely odd, in just the way that time-reversals of our world are odd: they will fail to accord with the "statistical macrolaws" of the actual world (e.g. Popper's asymmetry), at the very least in the time surrounding the miracle and almost certainly in the pre-miracle part too.[31] Obviously statistical macrolaws are crucial to our judgements of what is an apparent trace of what. But Lewis is very explicit (1986b, p. 57) that such statistically-based asymmetries should not be counted as matters of law, and he gives them no special weight in his account of similarity of worlds, so that he has no way to rule out of consideration worlds that violate them.[32] Lewis is offering an alternative to a statistical account of the direction of time, but there simply isn't the asymmetry he claims.

Lewis thinks that what underlies the (alleged) asymmetry of miracle-size and of traces is a non-statistical fact about the world: an asymmetry of overdetermination. For the laws of our world to determine whether (say) Fisk hit a home run in a certain baseball game in 1975 on the basis of prior information (say, information about what happened at a moment on the previous day), one needs a *vast* amount of prior information; perhaps (as Latham argues convincingly) information about the entire cross-section of the past light cone at that prior moment. Given how much is required of the one determinant at that prior moment, there is no room for any other. But to determine whether Fisk hit the home run on the basis of *future* information, Lewis thinks that much less will do, and that there will be a great many of these smaller determinants: this is plausible (he says) given that there are many independent traces of Fisk's home run (in newspaper archives and TV clips and reminiscences of Red Sox fans). Of course, none of these traces by itself is literally *sufficient* given the laws for Fisk's home run: there are lots of ways that any given one could be a result of fakery or whatever. But fakery leaves its own traces, and reflection on this is supposed to make plausible that

> very many simultaneous disjoint combinations of traces of any present fact are determinants thereof; there is no lawful way for the combination to have come

about in the absence of the fact. (p. 50)

In fact, however, while there is doubtless some sort of "near-determination" of Fisk's home run by many diverse facts about a given future time t_2, I don't think it is at all plausible that there is overdetermination by facts about t_2, for I don't think that there can be complete determination of Fisk's home run by any collection of facts about t_2 that doesn't include facts about the entire cross-section of the future light cone. Certainly there is no obvious way to *prove* determination by anything less than this.

What of the fallback idea of an asymmetry of *near*-determination? It is doubtless true that there is some sort of "near-determination" of Fisk's home run by many salient future facts, and that it is very unlikely that any *salient* facts about what happened before the home run (short of the state of an entire cross-section of the past light cone) give any "near-determination" of it. But there are plenty of other cases that go in the opposite direction: in a system isolated between t_1 and t_3, if it is in equilibrium at the intermediate time t_2 then that "nearly determines" its future state but not its past state. In any case, it should be clear that an analysis of "near-determination" would make it dependent on statistical regularities of the universe. This is to bring in an element that goes beyond those that Lewis considered.

My own view is that while it would be hard to find an acceptable statistical account of the directional asymmetry based on an asymmetry of near-determination, still bringing in statistical macrolaws in one way or another is the way we need to go, for there simply is no directional asymmetry independent of them. (Lewis informs me that this is now his view as well.) In my opinion, the best account of directional asymmetry is the one mentioned earlier in connection with the Spirtes, Glymour and Scheines theory of causal graphs: it's simply a fact that the variables we find salient tend to be probabilistically related in such a way that you can draw causal graphs among them in accordance with the Spirtes, Glymour and Scheines constraints; and (around here

anyway) what are causes on the causal graph criterion tend to precede their effects. This means that for systems too small for the statistical factors to show up, there is no "intrinsic" difference between cause and effect: it's simply that the temporal relation of the cause to the effect in such small systems is the same as the temporal relation between causes and effects among salient variables in larger systems (at least, in larger systems around here) as determined by the SGS theory. This is doubtless at odds with our pre-theoretic conceptions about cause and effect, but those pre-theoretic conceptions can not withstand what we have learned from physics. In this one regard at least, Russell was correct.[33]

Notes

1. It depends on which of the other concepts involved in the a law are treated as primitive and which as implicitly involving time: the latter but not the former may be allowed to shift in a transformation that reverses time. See Sklar [1974], pp. 364-8.

2. This requires a slight qualification, noticed by Earman [1986]: unless one assumes a finite upper bound on the velocity of propagation of forces (thereby ruling out Newton's law of gravitation), one must make some further assumptions about boundary conditions. But whatever qualifications are required must be made in the forward direction as well as the backward, so I don't think they much affect the basic point.

3. As I will note later, the asymmetry in statistical relations doesn't hold among *all* variables; the claim is only that it holds among variables *salient to us*. So the directionality in causation has a surprising anthropomorphic aspect as well as a surprising extrinsic aspect.

4. And the corresponding statements involving 'will' instead of 'must' are never *non-accidentally* correct: if they are exceptionless, it is only because the initial conditions required for an exception are never realized.

5. This is certainly implied by views (Mackie 1965, Bennett 1988) on which any part of a minimal sufficient condition for something happening at t is one of its causes; or by weaker views that imply this only when all parts of that minimal sufficient condition involve the same time, and that time is earlier than t.

6. I suspect that the difference between Bennett and Lewis is mostly terminological, over what they mean by 'event'. Another variant with little if any substantive difference from the Bennett view is that 'is a cause of' isn't a relation at all, but rather, is part of a sentential connective.

7. Of course, just as the natural causal conclusion to draw from the first set of data is undercut by the fuller data in the second set, so also the natural causal conclusion to draw from the fuller data in the second set could conceivably be undercut by still further data.

8. At the same time, if you have no independent knowledge of whether you have the gene, it is bad news to find yourself smoking, since this is evidence that you have the gene and have a good chance of getting cancer as a result of it despite your best efforts to prevent it by smoking.

9. I'm also not sure why the laws governing the causal fluid don't count as mere laws of association.

10. I'm not certain that the priority issue between singular claims and *system-restricted* general claims is ultimately a clear one.

11. Note that there are three different ways in which this is interest-relative: interests determine the appropriate contexts, they determine how significant a causal role is being claimed for smoking in those contexts, and they determine how high the probability of causation needs to be to count.

12. It might be thought that I shouldn't even really care about the chance of my smoking *causing* my cancer, but only about the conditional probability of my *getting* cancer on the assumption that I smoke and the conditional probability on the assumption that I don't smoke. If so, causation might seem irrelevant to effectiveness of strategy after all. I am sympathetic to the idea that what a person should be concerned about can be stated simply in terms of conditional probabilities; but this is defensible only on a special interpretation of the probabilities involved, and the required notion of probability is intimately bound up with the notion of causation.

The issues here have been more thoroughly discussed in the context of theories of rational decision rather than theories of effectiveness of strategy. A common claim (e.g. Skyrms 1980) has been that standard ("evidential") decision theory dictates the obviously unreasonable conclusion that even if one knows the Gene X story to be true one ought not smoke (if health is the only consideration), provided that one doesn't know whether one has Gene X. To avoid the conclusion, it is alleged that one must modify decision theory by building the notion of causation explicitly into one's rule of decision. But a number of authors have pointed out ways around the argument for replacing evidential decision

theory by a causal decision theory. Some (e.g. Horwich 1987) rely on the controversial assumption that the only way for Gene X to lead us to smoke is to produce in us an introspectible desire to smoke (this is called the "tickle defense"), but more recent authors avoid this: see for instance Price 1986 and 1991, Meek and Glymour 1994. Still, on the Price and Meek-Glymour accounts the agent's subjective probabilities are intimately bound up with her beliefs about causation, and some might argue that they need to be justified in terms of assumptions about causation (or about objective probability in some causally loaded sense). So the significance of avoiding explicit appeal to causation in the decision rule is controversial.

13. Of course, these dependences and independences can be masked by other common causes of X and Y.

14. Whether the light switch at one end of the hall is up may be independent of whether the light switch at the other end is up, even though the positions are highly correlated given that the light is off (they must be in opposite positions, barring a burned out bulb or melted wire), and almost perfectly correlated given that the light is on.

15. See for instance Ch. 1 of Maudlin 1994, or any of the essays in Cushing and McMullin 1989.

16. Or at least the fact that the causal ordering of the variables is backwards-discrete, i.e. that every node that has a non-immediate predecessor p has an immediate predecessor q of which p is a predecessor.

17. If our interest is in contrastive causation (Hitchcock 1993 and 1995), as it really should be, we need contrastive counterfactuals: the counterfactualist should say that for John's smoking cigarettes as opposed to his smoking cigars to have caused his cancer, it must be the case that if he had smoked cigars instead of cigarettes he would not have gotten cancer. The use of contrastive causal statements is especially important when we try to generalize to an account of causation for indeterministic contexts, but that lies outside the scope of this paper. (See Menzies 1989 for a discussion of one important issue about it.)

18. The distinction apparently collapses if, like Stalnaker, one takes the following to be a logical law:

(CEM) Either (if A were the case then B) or (if A were the case then not-B).

But even on Stalnaker's view the distinction remains, it just must be drawn differently. For in defending (CEM) against apparent counterexamples, Stalnaker concedes that it is sometimes indeterminate which disjunct holds. In that case, there is a distinction between it being determinate that if John hadn't smoked he wouldn't have gotten cancer and it not being determinate that if John hadn't smoked he would have gotten cancer, and we could use these "determinately counterfactuals" instead of plain counterfactuals in our theory of causation.

19. Even under determinism, there are various ways for John to have smoked and not to have smoked, ways not necessarily under John's powers to discriminate between, and whether cancer results is likely to depend on the way in which he smokes or fails to smoke. It is not out of the question to invoke a statistical probability measure over the ways of his not smoking, and perhaps the ways of his smoking as well; and it is worth distinguishing such a use of statistical probability from the use of any dynamic probability measure involved in fundamental indeterministic laws, because a theory might well invoke these different kinds of probability in different ways.

20. It should be clear from the discussion of the Russell-Latham problem that such a minimal sufficient condition for E will have to be very big: if it includes only information about a specific time prior to E then it will have to include some information about each point at that time that is in the past light cone of E.

21. Lewis notes that counterfactuals in which the time order of consequent to antecedent is the reverse of this tend to be given special syntactic markers (Lewis 1979, pp. 34-5.) Note though (contrary to what his discussion there suggests) that this syntactic "peculiarity" is not present in all counterfactuals that depend on backtracking arguments: it isn't present in the barometer counterfactual or the blue-eyes counterfactual, since the time order there is standard.

22. One approach that might do this would be based on the idea (mentioned earlier as a possibility) that system-specific general causal claims are prior to singular causal claims; if SGS-style causal models could be used in an account of the general causal claims, the direction of the arrows in such models could be used in an account of singular causation.

Another approach would be to separate the explanation of directionality from the account of counterfactuals proper, by building explicitly into the latter that the time of the antecedent must precede that of the consequent but then going on to explain why an account with this feature is more explanatorily useful than one with the reverse time order built in. (This is the analog for counterfactuals of something I suggested directly for causation in the next-to-last paragraph of Section 2.) Of course, such an approach would work only if the temporal restriction indirectly excludes common cause counterfactuals somehow.

23. If one wants events rather than facts in the cause or effect position, one can use counterfactuals about the occurrence of the events.

24. One might try to save the counterfactual dependence story by taking the relevant counterfactual dependence to be this: the fact that the window broke *in just the way it did* counterfactually depends on the fact that Joe threw the rock *in just the way he did*. But if we suppose that had Joe thrown his rock differently from the way he actually threw it, Pete would have stopped Joe's rock and thrown his own rock in the way he was poised to do in the story (i.e., in a way that makes its final velocity and time of arrival the same as on Joe's *actual* throw), there is no more a counterfactual dependence on Joe's manner of throwing than there is on his throwing.

25. (1) If Pete's rock passing a certain point at a certain time (with a certain velocity) would have been a different event from Joe's rock passing that point at that time (with that velocity), then for any point and time on the final segment of the trajectory of Joe's rock, the event E_1 of the rock passing that point at that time counterfactually depends on Joe's throw, and the window's breaking in the way it did counterfactually depends (in the non-backtacking sense) on E_1. (2) Even without relying on the special assumption about

event identity, we could rely on the fact that Pete can't have actually been in the path of Joe's rock (since if he had he would have been hit by it), so it would have taken him a certain amount of time to see that Joe wasn't throwing and to get into position to throw himself. So the window's shattering counterfactually depends on an event E_0 concerning the trajectory of a rock a microsecond before Pete would have had to throw, and of course E_0 depends on Joe's throw.

26. When C, E and A are true, saying that E counterfactually depends on C holding A fixed is simply to say that if A and not-C were the case then not-E would be the case. This is an ordinary non-backtracking counterfactual, though one whose antecedent involves multiple locations. (We need counterfactuals whose antecedents involve multiple locations for causal statements anyway, for cases where the cause involves what's going on at multiple locations.) Any apparent specialness in the particular multi-node counterfactuals here arises from the fact that in the cases of interest, A is an effect of C. But a restriction on backtracking is best formalized by cutting the causal inputs to each node in the cause; and this way of formalizing it gives the results we want in the multi-node counterfactuals that we need here.

27. Another kind of preemption ("trumping") is discussed in Schaffer 2000: the Major and the Sergeant shout the same order at the same time, and the Private follows the common order; but it seems as if the Major's order was the cause since had the two orders differed the Private would have obeyed the Major. It seems to me that any problem this raises for counterfactual theories is an artifact of using binary variables (Major either giving order C or giving no order, and similarly for Sergeant): a proper representation of the situation would allow the Major and the Sergeant each to give orders other than C, and this will show the counterfactual dependence on the Major's order but not the Sergeant's. (Indeed, the intuitive argument that the Major's order was the cause already appealed to the possibility of orders other than C.)

28. Even with the requirement that t_1 precedes t_2, the condition *may* fail to rule out the possibility that C and E are effects of a common cause; I doubt that there is a problem here when the laws are like those of the actual universe (alleged examples, such as those

on p. 45 of Bennett 1988, fail to take account of how big determining conditions must be: see my note 20), but there are imaginable laws for which the stipulation that t_1 precedes t_2 does not suffice.

29. Something like this point is made in Price 1991, though in my view he concedes too much to Lewis.

30. Some further alterations, e.g. in the direction of magnetic fields, are required also.

31. For instance, consider a world that obeys our laws except for a small miracle at $t + \in$ in 1973 and is just like our world after the miracle, and in which at t, Nixon is pushing the button to launch missiles at the Soviet Union. In such a miracle world, the apparent traces of Nixon not pressing the button would be illusory, which would require a failure of many macrolaws. Indeed, the failure of statistical macrolaws might be so drastic that a few minutes before t there is nothing recognizably a person that is continuous with the "Nixon" that pushed the button at t, and nothing recognizably a button in the surroundings either. Given that we tend to identify objects across possible worlds by similarity in their initial segments rather than in their final segments, it may not be entirely appropriate to describe the miracle-world in the way that I did, as one in which Nixon was pushing the button. I don't think this shows anything of interest about the direction of causation, it merely shows that we tend to use counterfactuals whose antecedents are temporally loaded.

32. Indeed, the problem of directionality would have lost a great deal of its initial punch if we had appealed at the start to statistical macrolaws such as the second law of thermodynamics. Such laws are evidently asymmetric and have a fundamentally different character in the forward direction than in the backward; indeed, even under the idealization that they are forward deterministic, they are not backward deterministic.

33. Thanks to Ned Block, Cian Dorr, Christopher Hitchcock, Paul Horwich, Lisa Warenski, and the editors for comments on previous versions.

Bibliography

Arntzenius, Frank 1993. "The Common Cause Principle", in Hull, Forbes and Okruhlik, *PSA 1992*, vol 2., pp. 227-37. East Lansing: Philosophy of Science Association.

Bennett, Jonathan 1984. "Counterfactuals and Temporal Direction", *Philosophical Review* 93: 57-91.

---------- 1988. *Events and their Names*. Indianapolis: Hackett.

Cartwright, Nancy 1979. "Causal Laws and Effective Strategies", *Nous* 13: 419-38.

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

Cushing, James, and Ernan McMullin 1989. *Philosophical Consequences of Quantum Theory*. Notre Dame: University of Notre Dame.

Davidson, Donald 1967. "Causal Relations", *Journal of Philosophy* 64: 691-703.

Earman, John 1976. "Causation: A Matter of Life and Death", *Journal of Philosophy* 73: 5-25.

---------- 1986. *A Primer on Determinism*. Dordrecht: Reidel

Elga, Adam (forthcoming). "Statistical Mechanics and the Asymmetry of Counterfactual Dependence". Philosophy of Science (Supplemental Issue PSA 2000).

Hitchcock, Christopher 1993. "A Generalized Probabilistic Theory of Causal Relevance", *Synthese* 97: 335-64.

---------- 1995. "The Mishap at Reichenbach Falls", *Philosophical Studies* 78: 257-91.

---------- (forthcoming). "The Intransitivity of Causation Revealed in Equations and Graphs".

Horwich, Paul 1987. *Asymmetries in Time*. Cambridge: MIT/Bradford

Hume, David 1748. *An Enquiry Concerning Human Understanding*. London.

Latham, Noa 1987. "Singular Causal Statements and Strict Deterministic Laws", *Pacific Philosophical Quarterly* 68: 29-43.

Lewis, David 1973. "Causation", *Journal of Philosophy* 70: 556-67. Reprinted in Lewis 1986, pp. 159-72.

---------- 1979. "Counterfactual Dependence and Times Arrow", *Nous* 13: 455-76. Reprinted in Lewis 1986, pp. 32-52; page references to latter.

---------- 1981. "Are We Free to Break the Laws?", *Theoria* 47: 113-21. Reprinted in Lewis 1986, pp. 291-8.

---------- 1986. *Philosophical Papers*, vol. II. Oxford: Oxford University.

---------- 1986a. Postscript to "Causation", in Lewis 1986, pp. 172-213.

---------- 1986b. Postscript to "Counterfactual Dependence and Time's Arrow", in Lewis 1986, pp.52-66.

---------- 1986c. "Events", in Lewis 1986, pp. 241-69.

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

Mackie, John 1965. "Causes and Conditions". *American Philosophical Quarterly* 2: 245-55.

Maudlin, Tim 1994. *Quantum Non-locality and Relativity*. Oxford: Blackwell.

McDermott, Michael 1995a. "Lewis on Causal Dependence", *Australasian Journal of Philosophy* 73: 129-139.

---------- 1995b. "Redundant Causation", *British Journal for the Philosophy of Science* 46: 523-44.

Meek, Christopher, and Clark Glymour 1994. "Conditioning and Intervening". *British Journal for the Philosophy of Science* 45: 1001-21.

Menzies, Peter 1989. "Probabilistic Causation and Causal Processes: A Critique of Lewis", *Philosophy of Science* 56: 642-63.

Pearl, Judea 2000. *Causality*. Cambridge: Cambridege University.

Price, Huw 1986. "Against Causal Decision Theory". *Synthese* 67: 195-212.

---------- 1991. "Agency and Probabilistic Causality", *British Journal for the Philosophy of Science* 42: 157-76.

---------- 1992. "Agency and Causal Asymmetry", *Mind* 101: 501-520.

Russell, Bertrand 1912-13. "On the Notion of Cause". *Proceedings of the Aristotelian Society* 13: 1-26.

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

Sklar, Lawrence 1974. *Space, Time and Spacetime*. Berkeley: University of California.

Skyrms, Brian 1980. *Causal Necessity*. New Haven: Yale University.

Spirtes, Peter, and Clark Glymour and Richard Scheines, 1993. *Causation, Prediction and Search*. New York: Springer-Verlag.

Vihvelin, Kadri 1991. “Freedom, Causation and Counterfactuals”. *Philosophical Studies* 64: 161-84.