Counterfactuals and temporal direction

Jonathan Bennett

[From *The Philosophical Review* 93, (1984), pp. 57–91]

1. Introduction

A forward counterfactual conditional is one whose consequent is about a later time than any the antecedent is about: 'If Stevenson had been President in 1953, the Viet Nam war would not have escalated in the 1960s.' The consequent of a backward conditional is about a time earlier than any that its antecedent is about: 'If Stevenson had been President in 1953, he would have won the election in 1952.' In this paper I shall offer a partial analysis which shows there to be no difficulty about allowing forward and backward conditionals in the same breath, as it were. This is in sharp contrast with a theory of David Lewis's[1] which provides for forward conditionals but not backward ones: when we conditionalize from times to earlier times, Lewis thinks, we adopt standards which are not those we use for forward conditionals and which he does not undertake to describe. My theory also contrasts with Frank Jackson's, in which each kind of conditional is given an official theoretic basis, but the bases are different and we are not allowed to combine backward and forward conditionals freely in a single operation. In short, Lewis's theory is asymmetrical, Jackson's is symmetrical but split down the middle, and mine is unified and symmetrical.

In expounding various theories, I shall assume that something like this is true:

> (P < Q) is true iff Q is true at all the P-worlds which are closest to the actual world.

What marks off one theory from another is its view about what makes a world a 'closest' P-world. Despite the superlative form of the word, I do not take it for granted that closeness is a matter of degree and that 'closest' means 'more close than any other.' Lewis thinks that closeness is a matter of degree, but Jackson doesn't, and I am not sure. More about that in section 15.

On some theories in which closeness is a matter of degree, there are no closest P-worlds, even if 'closest' means merely that none are closer. See Lewis's *Counterfactuals* 1.4 for a clear pointer to how all my uses of 'closest P-worlds' could be adjusted so as to take in those theories as well.

2. Closeness as ordinary similarity

Lewis's theory, presented in his book *Counterfactuals*, says that closeness is similarity, so that the truth of (P < Q)

[1] All references to works by their authors' names can be unpacked according to the bibliography. Lewis's theory was presented in his 1973 book and then amplified in important ways in his 1979 article.

depends purely on whether Q is true at all the P-worlds which are *most like* the actual world. Lewis explained that he was relying on our ordinary, everyday, intuitive notion of over-all similarity, the one we apply to faces and houses and towns and countries and—why not?—possible worlds (p. 92).

This theory does not mention physical law, however. In the twenty years between Chisholm's work on this topic and Stalnaker's, it had been generally assumed that the analysis of counterfactuals must bring in law somehow. For example, a Goodman-type theory would say something to the effect that

> (P < Q) is true iff Q is derivable by laws from P in conjunction with true propositions R which. . . ,

with the blank representing a problem which Goodman conceded he could not fully solve. That is equivalent to something of the form

> (P < Q) is true iff Q is true at all the causally possible[1] P-worlds at which. . . ,

with the same blank to be filled in when the complete analysis is known. Lewis, however, broke with this tradition and announced that the concept of similarity could go it alone. Insofar as laws have a special status in the analysis of counterfactuals, Lewis wrote, 'they need not have [it] by fiat' because they can be given it by argument (p. 73), and he offered such an argument, using a certain analysis of the concept of law as his premise. I am not persuaded by that argument, but no-one could doubt its conclusion that 'similarity of worlds in respect of their laws is an important respect of similarity, contributing weightily to overall similarity' (p. 75).

This falls short of requiring that closest P-worlds be causally possible, that is, that they perfectly obey the laws of the actual world. Lewis argues that similarity does not require that much lawfulness, and indeed that it sometimes positively requires breaches of law. Here is his argument for the view that most-similar P-worlds can sometimes be expected to contain miracles, that is, events which break the laws of the actual world.

Pretend that determinism is true: we should still have some true counterfactuals and some false ones; or so Lewis thinks, and I agree. Now, if P is false (at the actual world), then every causally possible P-world is unlike the actual world in respect of its whole history up to the time (T) to which P pertains. Any good statement of the determinist thesis will tell you that much, making it clear that any two worlds which are strictly determined by the same laws are unalike at time T only if they are unalike at every earlier time. So, if we want to evaluate (P < Q) where P is false, we must either accept as 'closest' some worlds which are unlike ours at all times earlier than T, or deem to be 'closest' some worlds which are just like ours up to about T and are then pushed off our course by a miracle—an event breaking some actual causal law.[2] Lewis took the latter option. 'Laws are very important', he wrote, 'but great masses of particular fact count for something too; and a localized variation is not the most serious sort of difference of law' (p. 75). This consideration, he concluded, 'seems plausible enough to deter me from decreeing' that the absence of a miracle must

[1] I use 'causally possible' to mean 'perfectly conforming to the causal laws of the actual world'.

[2] Throughout this paper, when I say that two worlds are alike at time T, I always mean that they are alike in respect of those propositions which are true at them and which are about T. Similarly for talk about what worlds are like 'through' certain periods of time, and all other locutions of mine which literally imply that worlds are objects which last through time and alter. Of course they don't, but the short-cut formulations which imply that they do are too convenient to forgo.

outweigh difference in particular fact through the whole of pre-T time.

It would deter me too, if my basis for deciding was mere offhand out-of-context judgment of comparative similarity—that is, if I had to decide by complying with this: 'Consider these two worlds and give me your impression—without asking why the question arises—as to which of them is more like the actual world.' Lewis's invocation of that 'familiar notion of comparative overall similarity' which 'somehow, we do have' suggests that that is his basis for deciding. We shall see that that is a misunderstanding, but in the meantime let us just stay with the fact that Lewis is willing to allow miracles at closest P-worlds, never mind why.

3. A difficulty for Lewis's theory

Lewis's reason for tolerating or requiring miracles at closest P-worlds, as a buffer against allowing 'closest' P-worlds to be unlike ours for vast stretches of the past, seems to discriminate against backward conditionals. But I cannot discuss that until I have presented an argument, independently discovered by Kit Fine (p. 452) and myself (p. 396), which seemed to us to show that in a quite unintended way Lewis's theory also discriminates against many forward conditionals.

We all think that (P < Q) can be true even where Q's truth would involve an enormous difference from what actually happened; but Lewis's theory seems to rule out all such conditionals. Let's suppose that at moment T in 1972 if Nixon had pressed a certain button a third World War would have occurred, the button being wired to other things so as to make that consequence inevitable. Now, a world at which WW3 occurs in 1972 is so unlike the actual world that we must conclude that there are more similar ones at which Nixon presses the button and miraculously the current dies in the wire, Nixon changes his mind, and WW3 does not occur. By Lewis's theory, then, the conditional 'If Nixon had pressed the button, WW3 would have ensued' seems to come out false, even in cases where we would all agree that it is true; and similarly for other conditionals with big-difference consequents.

The problem would be solved if miracles were banned at closest P-worlds. If the theory stipulated that if P itself is causally possible[1] then some causally possible P-world is closer to the actual world than is any P-world at which a miracle occurs, it would follow that some world where the button is pressed and events run their natural course is closer than any at which the button is pressed and the current miraculously dies in the wire. And so the conditional (button < catastrophe) would come out true, as desired. But Lewis had cut himself off from this treatment of the difficulty by his refusal to require in closest P-worlds any more lawlikeness than can be inferred from similarity.

4. Divergence and convergence miracles

Lewis responds to this in his paper 'Counterfactual Dependence and Time's Arrow'. From this we learn (though I misunderstood, until Lewis helped me) that we had exaggerated the scope of Lewis's reliance on the ordinary intuitive notion of similarity. He had intended that only as the taking of a firm plain-man stand in favor of the idea of similarity as such, against those who say that respects of similarity are so many and so incommensurable that judgments of the form 'x is more like z than y is' are never any use at all. He had not meant to commit himself to his theory's coming out right if the relevant similarity judgments were made only on the basis of one's off-hand, explicit, out-of-context opinions about whether world x is more like ours than world

[1] Until section 14 is reached, I shall assume that P is never in itself contrary to actual natural law.

y is. So he was not committed to accepting the judgments of comparative similarity that Fine and I relied on in our argument.

That does not trivialize Lewis's theory, as some have alleged. It is a substantive hypothesis that there is some relation of over-all similarity, reasonably so-called, that will do the job.

Someone who adopts that hypothesis should then test it by trying to describe the relevant similarity relation in some detail. He should try to see whether he can fill in the details in a manner which makes the theory assign to uncontroversial counterfactuals their agreed truth-values. That is what Lewis does in his 'Time's Arrow' paper. He identifies four respects of similarity that make up the over-all similarity relation, and assigns relative weights to them (or perhaps rank-orders them, in which case the theory would be that a dissimilarity in one respect outweighs any amount of similarity in lower-ranked respects). And it turns out that when this similarity relation is employed, the Nixon argument fails because its premises about the comparative similarity of worlds are not only *not implied* by Lewis's theory but—now that the theory has been amplified—are *inconsistent with* it. In a nutshell: when steering by Lewis's detailed similarity relation we get the result that closest P-worlds can contain miracles which launch the antecedent, not ones which intervene between antecedent and consequent.

In explaining how Lewis reaches this remarkable result, some mildly technical terms will be helpful, namely the phrases 'divergence miracle' and 'convergence miracle'. These probably explain themselves, but I shall play safe: if world w is like the actual world for some period ending at T, and unlike it for some period starting at T, and if the unlikeness is caused by an event occurring in w at T in conflict with the laws of the actual world, then that event is a *divergence miracle*. And the notion of a convergence miracle is the dual of that: if w is unlike the actual world for some period ending at T, and like it for some period starting at T, and if the likeness is caused by an event occurring in w at T in conflict with the laws the actual world, then that event is a *convergence miracle*. Thus, a divergence miracle pushes a world off the track of the actual world, while a convergence miracle puts a world onto the exact path of the actual world.

Lewis, then, offers a similarity relation which is specially tailored to produce the result that divergence miracles count less against similarity than convergence ones do. Put like that, it sounds a drastic ad hoc gerrymandering of the similarity relation. But Lewis does not put it like that; rather, he offers a similarity relation which distinguishes *small* miracles from *large* ones, and he argues that—from the standpoint of the actual world, at least—it takes a large miracle to create a convergence whereas a small one can suffice for a divergence. In Lewis's usage, a 'small miracle' is one involving only a very few breaches of the laws of the actual world, whereas a 'large miracle' involves a good many 'different sorts of violations of the laws' (not: violations of a good many different laws), or, as Lewis also says, 'a multitude of little miracles, spread out and diverse' (p. 471). It has turned out that some of the miracles Lewis wants to classify as large are not spread or scattered through large regions of space-time, and he tells me that he no longer stands by the 'spread out' bit of the account. We are left, then, with the contrast between small clusters of illegal events and large and various ones. When Lewis says that the latter count more for dissimilarity than the former do, he is on perfectly safe ground. Even someone who insisted on steering by the plain man's off-hand explicit out-of-context judgments would have to agree with that.

As for the thesis that small miracles can create divergences whereas it takes a large miracle to make worlds converge: Lewis writes persuasively about this. Clearly a small miracle can derail a world, and Lewis addresses himself to the rest of the thesis—that is, to the proposition that convergence requires a large miracle—in terms of the Nixon example. He points out that a mere dying of the current in the wire—a small miracle—does not put that world on track with the actual world. It averts World War Three, but does not restore complete parallelism: heat was generated which must go somewhere, specks of dust have been disturbed, light reflected from Nixon's thumb is already halfway to the moon, a few dozen molecules were knocked off the switch mechanism, and so on. To cram all those effects back into the box, and make the world in question perfectly like the actual world again, would require many distinct violations of actual laws, that is, would require a large miracle.

Lewis does not offer his *asymmetry lemma*—his thesis that convergence miracles cannot be small as divergence ones can be—as true from the perspective of just any world. If we consider 'a simple world inhabited by just one atom', he remarks, we shall 'doubtless conclude that convergence to this world takes no more of a varied. . . miracle than divergence from it' (p. 473). In section 12 below I shall mention one important upshot of this limitation in the scope of the asymmetry lemma.

The lemma is limited in another way which, though it does no harm to Lewis's position, needs to be understood. It depends on the fact that many and perhaps all divergence miracles are also convergence miracles, as I now explain.

Let W_0 be the actual world, supposing it to be deterministic in both temporal directions; let W_1 be the world which is exactly like W_0 right up to but not including time T, its laws being just like those of W_0 except that they permit Nixon's pressing the button at T. (I am pretending that the statement 'Nixon presses the button at T' is completely specific, so that my description of W_1 fits only one world.) Clearly, W_1 is unlike W_0 respect of T and all later times. Let W_2 be the world which is exactly like W_1 at T and thereafter, its laws being exactly those of W_0 . Thus, of the two worlds where Nixon presses the button at T, W_1 is like our world before T and its button-pushing is a miracle relative to our laws, whereas W_2 is unlike our world before T and its button-pushing is in perfect accordance with our laws. From T onwards, of course, W_1 and W_2 are indiscernible in all matters of particular fact. Now what should an inhabitant of W_2 think about the button-pushing at W_1? He should regard it as a miracle: it would not have happened in those circumstances in his world. But from his standpoint it is a small miracle, a mere re-routing of a few electrons in one brain. This is because it is, *ex hypothesi*, a small miracle relative to the laws of W_0; and so it must also be so relative to the laws of W_2, since these are the laws of W_0. Yet this event at W_1 which is (relative to W_2) a small miracle is also (relative to W_2) a convergence miracle. It is because and only because W_1 contains that event that from T onwards it is perfectly on track with W_2. That is what I said I would show, namely that an event which is a small miracle relative to a world can produce a convergence relative to that same world.

That refutes the asymmetry lemma as I have stated it. But that is mine, not Lewis's. All he claims is that it takes a large miracle to produce a convergence between the actual world (or one like it) and a plausible candidate for the title of closest P-world, where P is the antecedent of any counterfactual we are trying to evaluate; and I have no argument to show that any of those convergences could be produced by a small miracle. Notice also that Lewis's real topic is *re*convergence miracles, that is, events through which worlds which were

alike and then unalike become alike again. I can find no way of extending my argument to cover those.

5. The four respects of similarity

There is more to Lewis's similarity relation than just the relative rankings of large and small miracles. Here is the whole story. If x and y are worlds, y is closer to the actual world than x is, if

(1) *x contains a large miracle* and y does not; or

(2) Neither contains a large miracle, but *x has a smaller spatio-temporal region of perfect match with our world than y does*; or

(3) Neither contains a large miracle, and the regions of perfect match are equal, but *x contains a small miracle* and y does not; or

(4) Neither contains any miracle, and the regions of perfect match are equal, but *x is less similar to the actual world than y is*.

Presumably we have to apply **(4)**—at least until more work is done on it—on the basis of our off-hand judgments of comparative similarity. Anyway, **(4)** has an equivocal place in Lewis's theory. Although he thinks that when **(1)** through **(3)** are inapplicable, **(4)** sometimes yields the right answer, Lewis also thinks that sometimes it does not. There is evidence for this in the literature. For my purposes, however, all that matters is the claim that whenever any of **(1)** through **(3)** is involved, **(4)** cannot have any effect on the truth-value of the counterfactual in question.

This similarity relation smoothly handles the problem about conditionals like (button < catastrophe). Wanting this conditional to come out true, we wanted a world I'll call Catastrophe to be a closest world at which Nixon presses the button; but it had rivals, namely worlds where Nixon presses the button and a miracle intervenes and averts a World War. But now Lewis is distinguishing these rival worlds into two groups. There is the group containing worlds like Imperfect Convergence: this is the world Fine and I thought of, where Nixon presses the button, the current dies in the wire, and apart from that nature takes its course. This may be closer to the actual world at level **(4)** than Catastrophe is: the particular facts in those two worlds from 1972 onwards are conspicuously different. But this level-4 advantage has been purchased through a level-3 disadvantage, namely having a small miracle where Catastrophe has none. The vital fact is that Imperfect Convergence parts company with the actual world at the same moment as Catastrophe, and, like Catastrophe, never again becomes exactly like the actual world; so there is nothing to choose between them at level **(2)**, the level of extent of region of perfect match. Typical of the other group of rival worlds is Perfect Convergence. In this, Nixon presses the button, and a number of events occur which make the situation exactly as though he had never done so. This world is exactly like the actual world throughout its entire history except for a second or two during and just after the pressing of the button; it thus has an enormous advantage over Catastrophe at level **(2)**; but this has been bought at the price of a disadvantage at level **(1)**, since Perfect Convergence contains a large miracle whereas Catastrophe does not. So both rivals fail, Catastrophe remains the closest, and the conditional (button < catastrophe) comes out as true.

I have mentioned one respect in which the similarity relation that produces this brilliant result is intuitively natural, matching our casual off-hand judgments about similarity, namely its putting **(1)** large miracles above **(3)** small ones. I now add that Lewis tries also to make it seem natural to put **(2)** extent of perfect match higher than **(4)** degree of imperfect similarity. He does this by making **(4)** look trivial in the long run, suggesting that even the tiniest differences between

worlds amplify as time goes on. In Imperfect Convergence, for example, one of those specks of dust will get in someone's eye, one thing will lead to another, and within a thousand or a million years that world will differ from the actual world as much as Catastrophe does. This thesis about amplification is not required for the theory but does help to make it look better.

6. Why are miracles still tolerated?

Still, there can be no question of the theory's relying on nothing but intuitive off-hand similarity judgments. Granted that **(1)** intuitively has to come above **(3)**, it is less obvious that **(2)** must outrank **(4)**, and the interleaving of **(1–3)** with **(2–4)** is simply stipulated. The theory has to say that this novel **(1–2–3–4)** item just is the relation which so defines '. . . is closer to the actual world than. . . is' as to make it the case that (P < Q) is true if and only if Q is true at the closest P-worlds.

What is the case for tying counterfactuals to this similarity relation rather than to one which bans all miracles at closest P-worlds? Lewis's remark in *Counterfactuals* that if laws have a special status 'they need not have it by fiat' suggests that he won't prohibit miracles because that would require him to bring in the concept of miracle, and thus of law, into the premises of the theory whereas he thinks it need not come in except derivatively. But Lewis assures me that this is taking his remark more strongly than he intended it; and anyway even if he had once aimed to analyze counterfactuals without help from the concept of law he is not doing so now; for the concept of miracle, and thus of law, is busily at work in the 'Time's Arrow' delineation of the crucial similarity relation. So, I repeat, why not ban all miracles from closest P-worlds?

The only other apparent answer in *Counterfactuals* is an appeal to plausibility: Lewis finds it 'plausible' to suppose that a world with a history like ours up to T and a miracle at T is more like ours than is a world with no miracle and T and a history unlike ours for all earlier times. I find that plausible too, if I steer by my ordinary intuitive off-hand impressions of comparative similarity. But Lewis has since warned us that those are an unreliable guide to the similarity judgments that are needed in evaluating counterfactuals; and I can find no other reading of his remark about plausibility that makes it a force to be reckoned with.

In short, I cannot find anywhere in *Counterfactuals* any cogent basis for refusing to ban all miracles at closest P-worlds.

If all miracles were prohibited, a closest P-world might be unlike the actual world in respect of all times earlier than T—but why not? It is not enough to say that that is ruled out by rank **(2)**—extent of perfect match—in Lewis's new similarity relation; for that is what I am challenging. Why should rank **(2)** have any part in the account? It did not help to solve the problem about (button < catastrophe), but merely affected the shape that the problem took. It was because Lewis put **(2)** extent of perfect match above **(4)** degree of imperfect similarity that he had to put **(1)** big miracles above **(3)** small ones. Abolish both discriminations, putting a fused **(1–3)** above a fused **(2–4)**, and you get a theory according to which a closest P-world can contain no miracles, and as between two unmiraculous worlds the one which is more over-all similar to the actual world is the closer. This still preserves (button < catastrophe), by banning any miracle which might intervene between antecedent and consequent; and it does not conflict with anything I have so far reported Lewis as saying.

7. Objections to miracles

The need for reasons is a pressing one. Lewis cannot simply say that his position is plausible enough to be acceptable in the absence of objections to it; for there are objections. They might be overcome, but only by a hard push going his way.

(i) I find it objectionable that the new theory, like the old one, makes false most counterfactuals to the effect that if P were the case at T then Q would have been the case earlier, or—to use the terminology which I prefer and which Lewis also finds more 'natural'—if P were the case at T then Q would *have to* have been the case earlier. It seems to me just plainly true that if the actual world is deterministic then if a certain pebble had rolled at a moment when in fact it did not roll, the entire previous history of the world would have had to be different. If you don't agree with me about that, then I invite you to agree that that counterfactual is plausible enough to merit retention unless there are strong reasons to reject it.

That will be rejected by those who think that if (P < Q) is true then—to put it in short-hand—there must be a causal flow from P to Q; for if that were right then I would be implying that the rolling of the pebble can affect past history. But why accept that premise? We plainly do sometimes assert counterfactuals which run against the causal flow; for example, saying that if the die had fallen six uppermost it would (have to) have been thrown differently.[1] What is needed, then, is an independent reason for keeping such counterfactuals at a distance from the forward ones which are everybody's primary concern: if not for writing them off as unworthy of philosophical respect, then at least for segregating them, not allowing them to be handled by the same standards as are used in evaluating forward conditionals.

(ii) Pollock has noted another problem, arising from Lewis's stress on **(2)** extent of perfect match.[2] Lewis's position implies that if a given world is a closest P-world, and involves a divergence miracle, the miracle must be small (because of rank **(3)**) but also late (because of rank **(2)**. This is not an inconsistency, but it creates a tension: the earlier you have the miracle, the smaller it can be; but, on the other hand, if you delay it until later you increase the extent of the temporal region of perfect match with the actual world. Some kind of trade-off is needed, then. Pollock's problem is as follows. I left my coat in the cloakroom last night, and it was still there at noon today. Conditionals starting 'If (P) my coat had been gone by noon today, then. . . ' take us to closest P-worlds. Now, suppose that for the coat to have gone would require a tiny miraculous event in any one of several different brains, belonging some to people who were near my coat last night and some who were near to it this morning. Lewis has to favor this morning over last night, so as to choose a world which diverges from ours as late as possible; and so he should regard as true the conditional 'If my coat had been gone by noon today, it would have to have been taken at some time this morning.' That seems to be just wrong, since a theft last night is equally possible and probable. A good theory may assign truth-values where intuition is silent; but the conditional favoring a morning theft is a creature of theory which conflicts with intuition.

[1] David Sanford has pointed out to me how dangerous it is to say that there is a *dependence* of consequent on antecedent in every true counterfactual, even when embedded in the phrase 'counterfactual dependence.' It suggests that whether the consequent obtains depends upon whether the antecedent obtains, and for counterfactuals running against the causal flow that is not the case. Lewis does not argue from this false suggestion in his terminology, but it would be better if the suggestion were avoided altogether.

[2] Reported in Donald Nute, *Topics in Conditional Logic* (Dordrecht, 1980), p. 104.

8. The Downing scare-story

There are, then, objections to Lewis's tolerance of divergence miracles and his associated emphasis on the extent of the period of perfect match. But he argues that a viable theory must have those features or something like them. In my review of *Counterfactuals* I applauded that aspect of the first theory, and offered two reasons for it. (I wrongly thought I was adding to a couple of reasons Lewis had already given—that we can and should keep 'law' and 'miracle' out of the premises of the theory, and that we should trust our off-hand judgment that a long dissimilar history makes for more dissimilarity than does one miracle.) They purported to show that we must keep backward counterfactuals at arm's length, not allowing them to mix in with more usual forward ones; and so they promised to clear the way for the position that in evaluating forward conditions we may be able to count some worlds where divergence miracles occur as being among the closest P-worlds. In his 'Time's Arrow' paper Lewis accepted both of these reasons (pp. 456, 469).

The more important of the two was first thought up by P. B. Downing (pp. 125–126) more than twenty years ago. Here is a version of Downing's scare story:

> Mr. D'Arcy and Elizabeth quarreled yesterday, and she is still very angry. We conclude that if he asked her for a favour today, she would not grant it. But wait: Mr. D'Arcy is a proud man. He never would ask for a favour after such a quarrel; if he were to ask her for a favour today, there would have to have been no quarrel yesterday. In that case, Elizabeth would be her usual generous self. So if Mr. D'Arcy asked Elizabeth for a favour today, she would grant it after all.

From this we are to infer that it is dangerous to combine forward and backward counterfactuals in a single operation. That is Lewis's main argument in his 'Time's Arrow' paper for continuing to allow divergence miracles at closest P-worlds: if he didn't do so, those worlds might be richly unlike the actual world in respect of all of pre-antecedent time, and so they would justify infinitely many backward counterfactuals which would interact fatally with the forward ones.

All honor to Downing for noticing that there is a question about backward counterfactuals; but no credit to any of us for being taken in by this quite ungrounded scare story. We are invited to consider a pair of conditionals of the form 'If Mr. D'Arcy had asked Elizabeth for a favor at T,. . . . '. On any reasonable theory, we must evaluate these by looking for the closest worlds at which Mr. D'Arcy does that, and to know which worlds those are we must see what the actual world is like at T. The Downing story assumes that we shall evaluate the forward conditional by looking at Elizabeth's anger while ignoring Mr. D'Arcy's pride, and evaluate the backward one by taking account of his pride but not her anger. Of course this will get us into trouble! But the trouble has nothing to do with combining the two temporal directions: it comes entirely from indecision or inconsistency regarding what facts about the actual world are to be taken into account.

A similar pattern is shown by the other version of the Downing argument in the literature.[1] In circumstances where it is right to say that if I jumped out of that window I would be killed, that conditional is said to be threatened unless we keep backward conditionals at arm's length, because: given my prudent character, if I jumped out of the window I would have previously arranged for a safety net to be placed underneath, and so if I jumped there would be a net and

[1] Jackson, p. 9. It is only fair to report that Jackson does not endorse the argument as clearly and forthrightly as Lewis does.

I wouldn't be killed after all. This is the same muddle as before: different standards for closeness to the actual world are arbitrarily associated with different temporal directions. In fact, the different standards can run us into contradictions even if we stay with forward conditionals. If we are allowed sometimes to ignore the absence of a safety net, then let's instead ignore how high the window is; then we can get 'If I jumped out of that window I would [or at least: might] not be killed', thus getting a conflicting pair of forward conditionals. The trouble has nothing at all to do with temporal direction.

No doubt we do have somewhat different ways of looking for the closest P-world, depending on aspects of the context—for example, depending on whether our counterfactual has come up in discussion of individual psychology or group dynamics. As a mere illustration of that fact, the Downing story succeeds. But Downing and I have used it on the assumption that of two different ways of determining the closest 'D'Arcy asks' worlds, one is right only if the conditional runs forward in time while the other is right only if it runs backwards; and there is no warrant for that. I believe that Lewis belongs to our guilty group, though his discussion of this matter (in the 'Time's Arrow' paper, pp. 456ff) is not quite explicit about it. Lewis says that we have different ways of 'resolving the vagueness' of the antecedent of a counterfactual; he speaks of our 'standard resolution' of vagueness, used when evaluating forward conditionals, and says that if someone propounds a backwards conditional his listeners, if they are co-operative, 'will switch' to a 'special resolution that gives him a chance to be right'. I cannot understand this if it does not rest on the belief, presumably drawn from the Downing story, that no single 'resolution of vagueness' will allow us to go in both directions from the same antecedent. That is what I am challenging.

Perhaps I have misunderstood this passage of Lewis's. Perhaps he does not mean it as an argument for what he calls the temporal 'asymmetry of counterfactual dependence', that is, for tolerating divergence miracles at closest worlds so as to dam the torrent of backward conditionals. If so, I cannot see that he has any substantial argument for that important aspect of his position. For all he is left with is the argument of mine which I added to Downing's in my review, and which Lewis endorses in passing on his page 469. That argument is thoroughly bad, however, as I shall explain at the end of section 9 below.

I have been challenged to explain why the Downing examples are plausible. Well, perhaps that is explained by some fact about how we usually handle backward conditionals about human action. In accepting the backward conditional about Mr. D'Arcy, we start not with the world, but only with the Mr. D'Arcy, which is closest to the actual one today; we unroll the story of that Mr. D'Arcy back for a day or two and then, finding within him no traces of a quarrel, we conclude that there was no quarrel; and only then do we enlarge the frame so as to include Elizabeth in it, running her story forward again from a quarrel-free yesterday to a sweet-tempered today. Perhaps this is typical of our approach to backward conditionals about human conduct, though I doubt it; perhaps there is even some rationale for it, though I doubt that even more; but whatever is going on here could not possibly be typical of the evaluation of backward counterfactuals generally.

Really, I don't think that anything systematic is going on when people accept Downing stories. The answer to 'Why are they plausible?' is 'They are not'. Our falling for them was merely careless.

9. A unified symmetrical theory

I now propose a partial theory of counterfactual conditionals which is perfectly symmetrical with respect to temporal direction. I state it only for conditionals whose antecedent is about a particular moment or period of time T; it can be extended to capture the others—partly by methods described by Jackson (pp. 12–14), partly by quantifying over times, and so on. I shall not discuss any of that.

My theory starts with the idea of a T-closest P-world, meaning a P-world which is closest to the actual world at time T. In expounding and illustrating this, I shall assume that T-closest worlds will be very similar to one another at T, but that is not part of the theory. My theory is only partial because it does not say what T-closeness consists in; I merely assume that if it doesn't consist in T -similarity then it at least implies it.

I offer, as a first approximation to the theory (which I shall modify at the end of section 10 below), the proposal that (P < Q) is true if and only if Q is true at all the T-closest causally possible P-worlds—where T is the time to which P pertains. According to that, you learn whether a counterfactual is true by finding the T-closest antecedent worlds which obey the laws of the actual world, and discovering whether the consequent is true at those worlds. Or, in the language of world stages, you find the closest T-world-stages, unroll the rest of those worlds—for all times earlier and later than T—in accordance with the laws of the actual world, and see whether any of them contain Q. If all of them do, (P < Q) is true; otherwise false. There is here no bias in favor of conditionals running from earlier times to later, no provision for any miracles, and not the remotest hint of a threat from the Downing scare story. In the D'Arcy example, if we start with the T-closest world where D'Arcy asks for a favor, it will probably be one where D'Arcy doesn't think there has been a quarrel and Elizabeth does think there has; working back from that in accordance with actual laws, we'll presumably encounter a failure of memory on his part or a false memory on hers; but we certainly shan't get a quarrel-free yesterday and Elizabeth generously disposed today. My theory cannot possibly lead to contradictions: it starts from something causally possible, and adds only what is derivable from it through actual causal laws; so there cannot be contradictions or even miracles. (A reminder: counterfactuals with causally impossible antecedents are being set aside until section 14.)

There is no impediment in this to conditionals which go in both temporal directions at once; that is, ones of the form (P < (Q & R)), where Q pertains to a time earlier than the time (T) that P is about, and R pertains to a time later than T. What must not be done is to bring in facts about the actual world at times other than T. If T is the present, then we must put away our history books and crystal balls, using only our eyes and our capacity for causal inference in both temporal directions. That is my cue to own up to something. As well as the Downing scare story, I offered Lewis an argument of my own in favor of allowing divergence miracles (p. 391). The argument as written is a murky affair, but I remember what I meant: I was assuming that in evaluating backward counterfactuals we may freely work back in accordance with causal laws while *also* freely consulting the history books of the actual world. In predicting that this procedure would lead to trouble, I was right. My error was in treating this muddle as essential to the evaluation of backward conditionals. I am sorry that I ever fell into this confusion, and misled Lewis into thinking that Downing and I had produced two solid obstacles to backward counterfactuals, when really there are none.

10. History books and crystal balls

Lewis has maintained that in handling normal forward counterfactuals we do freely plunder the history books of the actual world ('Time's Arrow', p. 456). We accept conditionals in which the consequent is based partly on what is true after T at the T-closest P-world and partly on what is true before T at the actual world. For example, we may accept 'If (P) Smith had told all he knew to the police today, he'd have got his revenge for what Jones did to him yesterday.' My theory implies that we ought not to accept this unless we are satisfied that Jones wronged Smith yesterday at the today-closest P-world; so if Lewis is right then my theory condemns something we do freely and often; and a theory which imputes so much rash carelessness is presumably false. So the argument goes.

Well, how freely do we help ourselves to actual history books in forward conditionals? Not as freely as Jackson's theory implies. Jackson has accepted this point of Lewis's in its fullest possible strength, and has offered a two-part theory which says in effect that a forward (backward) conditional (P < Q) is true if and only if Q is true at the T-closest P-world which obeys the laws of the actual world after (before) T and is exactly like the actual world before (after) T (pp. 9, 11–12). That is, for forward conditionals we go by similarity up to the antecedent time and law thereafter; for backward ones we go by similarity back to the antecedent time and law theretofore. I have described this theory as (unlike Lewis's) symmetrical, and (unlike mine) fragmented, because it provides for conditionals in both temporal directions but not on the basis of a single standard of world-closeness.

Here is evidence against Jackson's theory of forward conditionals. 'If Adlai Stevenson had been President of the U.S.A. in (T) February 1953 then at his death the obituaries would have spoken of him as 'Eisenhower's liberal successor.' That is absurd: if Stevenson were President in 1953, he would have won the election in November 1952 and would have beaten Eisenhower rather than succeeding him. But not according to Jackson's theory. It evaluates the given conditional by looking at P-worlds which are exactly like the actual world up to T, and at those worlds Eisenhower is indeed President in January 1953 and is succeeded by Stevenson before the end of February.

The other half of the theory—which bases backward conditionals on pre-T law and lavish helpings of post-T actual fact—is also vividly in trouble. Readers should not have much trouble in devising examples which illustrate this.

That refutes both halves of Jackson's theory (it also refutes Wayne Davis's; see his p. 554). It also suggests that my theory deserves a rehearing. When in our forward conditionals we help ourselves to pre-T actual fact, do we really do this with a careless confidence which my theory would condemn? I see no evidence that we do. I suggest that our procedure is as follows. In accepting such conditionals, we do not minutely examine whether we are right in particular cases—for example, whether at the today-closest Smith-tells-all worlds Jones wronged Smith yesterday—but rather assume that this is so unless we see obvious reasons for suspecting that it isn't. This sensible attitude is exactly the one we would adopt if my theory were correct.

There is a vagueness in that account of our procedure, however. What if Jones wronged Smith at some of the relevant worlds but not at others? That would imply that Jones's deed had not left its mark on the actual world today; for any tracks it left in the here and now would appear in *all* the today-closest Smith-tells-all worlds—unless Jones's deed was incompatible with Smith's telling all he knew, in which case the tracks would appear in *none* of those worlds. But never mind that implausibility. What matters is the abstract

possibility of a proposition's being true at the actual world and at some but not all of the T-closest P-worlds. There are countless instances of this: although there were heuristic reasons for working with the assumption that the actual world is deterministic, we want a theory which will also let counterfactuals be evaluated at the actual world even if it is not deterministic; and a nondetermined world will be a rich source of cases of the sort under discussion. Here is an easier example. At T_1 I bet that when the coin is tossed at T_2 it will come up heads; and in the upshot it does just that; but this is a purely chance event, with no causally sufficient prior conditions. Now consider the conditional 'If I had bet on tails at T_1 I would have lost.' Everyone I have polled is inclined to say that that conditional is true, despite the fact that at some of the T_1-closest 'I bet on tails' worlds the coin comes up heads at T_2. (Why does it come up heads at some of those worlds? Because, since the fall of the coin had no causally sufficient prior conditions, every 'tails' world is indistinguishable, in respect of its state at T_1, from some 'heads' world.) If I am to respect these judgments I must modify my theory, replacing the clause '$\mathcal{Q}$ is true at all the T-closest causally possible P-worlds' by something like this: '$\mathcal{Q}$ is equivalent to some conjunction (R & S) such that R is true at all the T-closest causally possible worlds and S is true at some of those worlds and also at the actual world.' This is to be understood as being true if $\mathcal{Q}$ itself is true at every T-closest P-world, and if $\mathcal{Q}$ itself is true at some of those worlds and at the actual world.[1]

From now on I shall write as though I were still offering the simpler view that (P < $\mathcal{Q}$) is true just in case $\mathcal{Q}$ is true at all the T-closest causally possible P-worlds. This pretence merely enables me to avoid cumbersome formulations.

11. Lewis's theory versus mine

The facts which condemn Jackson's theory do not condemn Lewis's. Lewis can say that the 'revenge' conditional is true because the best candidate for the role of 'small late miracle leading to Smith's telling all to the police' is a miraculous brain event today; this post-dates Jones's wronging of Smith yesterday, and so leaves it untouched. On the other hand, the 'Stevenson' conditional is false because a *small* miracle leading to Stevenson's being President in February 1953 would have to occur before the election, starting a train of events which would include Eisenhower's losing.

Lewis's theory, then, is no worse off than mine with regard to this matter, but no better off either. I don't think there could be a case which was plausibly handled by Lewis's theory and not by mine, or vice versa. At this point in the battlefield we have a stand-off.

Looked at along the whole line of confrontation, however, there is much to recommend my theory over Lewis's. Mine is simpler, it can explain as much as his can, and it affords a domesticated welcome to backward conditionals which his theory keeps at arm's length. In discussing the Downing story, Lewis allows that the backward view of things—if he requested a favor today there would have been no quarrel yesterday—is permissible; but he contrasts it with forward conditionalizing as 'special' to 'usual' or 'standard', and says that when we are in the 'special' frame of mind 'we very easily slip back into our usual sort of counterfactual reasoning' (pp. 456ff). I cannot find, in our everyday handling of conditionals, evidence of any such deep division. Of course that would not matter if the division were needed; that is, if we could not safely conditionalize in both temporal directions

[1] That position is endorsed in my paper 'Even If', *Linguistics and Philosophy* 5 (1982): pp. 403–418, at pp. 414–417. I am indebted to Richmond Thomason for alerting me to its relevance to the present topic.

at once. But what reason have we been given to believe that? Only the empty Downing scare story and the muddled Bennett addition to it.

[Added in 2011: In my *A Philosophical Guide to Conditionals* (Oxford University Press 2003), section 80, I explain why the theory of mine that I have been praising here is certainly false. I still stand by most of the content of the present paper.]

12. Time's arrow

However, as well as those unsuccessful arguments for allowing miracles at closest worlds, Lewis also has a motivation. Developing an idea of Downing's,[1] he boldly attempts to extract from his theory of counterfactuals a basis for the notion of 'time's arrow', that is, our sense that the future is open, the past closed. He interprets this as meaning that the future depends counterfactually on the present in a way the past does not:

> 'We can bring it about that the future is the way it actually will be, rather than any of the other ways it would have been if we acted differently in the present. The future depends counterfactually on the present. It depends, partly, on what we do now. Something we ordinarily cannot do is to bring it about that the past is the way it actually was, rather than some other way it would have been if we had acted differently in the present. The past does not at all depend on what we do now. It is counterfactually independent of the present' (pp. 461–462, quoted with omissions).

Lewis is here offering his theory of counterfactuals, which accords truth to many forward counterfactuals and few backward ones, as explaining both the meaning and the truth of the common idea that the future is open, the past closed. (When he says that we 'ordinarily' cannot affect the past, he is leaving room for the possibility of temporally backwards causation. That would make time less arrowed than most of us believe it to be, but it is irrelevant to Lewis's and my present concerns.) Neither Jackson's symmetrical theory nor mine could possibly explain the direction of time's arrow.

Lewis's explanation gets extra power from the fact that his analysis derives a temporally asymmetrical output from an input which says nothing about temporal direction. With the aid of his thesis that divergence miracles can be small but re-convergence miracles are larger (see section 4 above), Lewis can derive from a premise about *small/large* a conclusion about *forwards/backwards*, through the mediation of the fact that divergence differs from convergence as forward differs from backward. The conclusion, of course, is that there are many counterfactuals running forward in time and very few running backward.

(Notice that Lewis's explanation makes it a contingent truth that time has an arrow; that is, that the past is closed and the future open. He points out that it does not hold at certain very simple worlds; and, as I added late in section 4 above, it does not hold either at complex worlds where small miracles can produce convergences on close other worlds. This contingency will be surprising to some, but it may well be right.)

Lewis allows that we can have as many backward counterfactuals as we like, if we adopt criteria of world-closeness appropriate to them. He does not say what criteria would be suitable; but we do know that the adoption of them is a non-'standard' kind of procedure which we 'very easily' slip out of. I shall now speak of 'standard closeness'

[1] 'The past is "inviolable" in that it cannot be true that if something happened now some past event would be different from what it would otherwise have been.' Downing, p. 136.

and 'standard truth', meaning closeness and truth according to the **(1–2–3–4)** similarity criterion of Lewis's which I expounded in section 5 above. So I can put his view about time's arrow—in a first approximation—by saying that standardly true counterfactuals run forward in time, not backwards.

13. Why that explanation fails

Never backward? Well, hardly ever! And there's the rub: Lewis must allow that some backward counterfactuals are standardly true, namely ones about the divergence miracle and the transition from that to the state of affairs mentioned in the antecedent. If Stevenson were President in February 1953, he would have been elected in November 1952: Lewis cannot prevent that from coming out as standardly true in his theory; for example, by saying that at the closest relevant worlds Stevenson was elected Vice-President under Eisenhower who then died. So he has a standardly true backward counterfactual. Indeed, he must allow some which stretch far back in time: although divergence miracles must be as late as possible, they must not be left so late that they need to be large; and sometimes they can be kept small only by being quite early, leaving plenty of time for the small miracle to cause the truth of the big-difference antecedent, and thus plenty of pre-antecedent time which depends counterfactually on the antecedent. For example, we are willing in principle to say things like 'If in 1933 there had been twice as many Jews in Germany as there actually were, then. . . '; but of the worlds where that is the case, those which are closest by Lewis's standards must have diverged from the actual world many years before 1933, for a late divergence in this case would require a big miracle, which Lewis can't allow at any price. So the sluices are open to floods of backward counterfactuals. There seems to be nothing left of the supposed foundation for time's arrow.

Lewis mentions this trouble, but seems to take it lightly. In describing the transition period from the miracle up to time T, he says:

> That is not to say, however, that the immediate past depends on the present in any very definite way. There may be a variety of ways the transition might go, hence there may be no true counterfactuals that say in any detail how the immediate past would be if the present were different. I hope not, since if there were a definite and detailed dependence, it would be hard for me to say why some of this dependence should not be interpreted—wrongly, of course—as backward causation over short periods of time [and thus, presumably, as an openness of the recent past] in cases that are not at all extraordinary' (p. 463).

The conditionals in question cannot be dumped in the non-standard bin: they are standardly true; that is, they concern what is the case at those standardly closest P-worlds where the P state of affairs flows from a recent miracle through a short transition period. So Lewis must neutralize them in some other way, and he hopes to do this by supposing that their consequents are not 'definite and detailed'. Well, I don't see why they aren't. For example, any late, small miracle leading to Stevenson's being President in February 1953 would surely lead to his being elected in November 1952—isn't that definite and detailed enough?

I am sure that Lewis did not say quite what he meant. Noting that he has suddenly switched from 'dependence' to 'causation', and remembering his theory about the latter,[1] one can hardly doubt that he meant to express the hope

[1] David K. Lewis, 'Causation', *Journal of Philosophy* 70 (1973): pp. 556–567.

that (except in 'extraordinary' cases), no standardly true conditional implies, for any event e_1 and subsequent event e_2, that if e_2 had not occurred then e_1 would not have occurred. He could defend that against any changes I can ring on my Stevenson example. If Stevenson's noninauguration had not occurred, then... but Lewis can plausibly deny that there is any such *event* as Stevenson's noninauguration. If Eisenhower's inauguration had not occurred, then... what? Then Stevenson's defeat (or Eisenhower's victory) would not have occurred? That has the right form, but Lewis can reasonably conjecture that it is false, because worlds where Eisenhower loses are no closer than ones where he wins and then dies in December leaving Vice-President Nixon to be inaugurated.

This, though, allows Lewis to explain only a specifically *event-causation* version of time's arrow, which means that he is not explaining or justifying my belief in time's arrow or, I suspect, yours. I think that the past is closed and the future open in respect of the states of affairs that obtain in them, including indefinite and undetailed ones; or, to use Lewis's own words with my italics, I think that 'the past does not *at all* depend on what we do now'. According to my version of the time's arrow assumption—which I am sure is the usual one—no fact about the world's state at any time before T depended on the fact that a certain apple fell from my tree precisely at T. But Lewis must allow that there may be standardly true counterfactuals running from the negation of the fact back into a nonactual past—the least informative of them being 'If that apple had not fallen at T then the world's previous state would have been somewhat different from what it actually was.' I cannot parlay that into anything of the form 'If e_2 had not occurred, e_1 would not have occurred earlier', but what of that?

I cannot explain time's arrow; I wish I could. But that does not incline me to settle for a theory which explains part of my time's arrow belief only if the part it does not explain—the part not expressible in terms of event causation—is downright false.

I think, then, that Lewis's theory about time's arrow does not succeed. So as well as having no good arguments for allowing miracles at closest P-worlds, he also has no sound motivation for allowing them.

It is interesting—I note in passing—that Lewis expresses the idea about time's arrow in terms of how past and future relate to 'what *we do* now' rather than to 'what *happens* now'. That might suggest that backward counterfactuals are blocked by the existence of radical freedom: if I was not causally determined to do A rather than B, then no backward conditionals of the form 'If I had done B...' are true. And even if I were causally determined to do A, it will often be plausible to suppose that my doing B would have followed from a very small extremely recent miracle, thus providing almost no room for a backward conditional. But that cannot help Lewis, for our ordinary ideas about time's arrow are not confined to the past's invulnerability to present human action; nor, I believe, does Lewis think that they are.

14. Counterlegals

So far, I concede no advantages to Lewis's second theory, and claim several for mine. A further feature of mine which might be thought disadvantageous is really, I shall argue, one of its merits. It has to do with counterlegal conditionals—ones whose antecedents conflict with causal laws, perhaps by saying false things about them. Until now I have set these aside, stipulating that all my conditionals have causally possible antecedents, and one might wonder whether my theory can be extended to take them in.

There is no trouble with the weak counterlegal which merely says that if P were the case then something contrary to causal law would be the case, or (P < a miracle occurs). That is equivalent to saying that no P-world is causally possible, or that there is a miracle at every P-world; and because it generalizes over all the P-worlds, needing no closeness relation or other device for selecting from amongst them, it creates no problems. The interesting and troublesome counterlegals are those whose consequents are more specific than that, saying that if causally impossible P were the case then Q would be the case, where Q is true at some causally impossible worlds but not at all. If counterlegals of that type are to be sorted into true and false, we need a suitable way of selecting from among the causally impossible worlds. How is this to be done?

Well, I have made the truth of (P < Q) depend on whether Q is true at the T-closest P-worlds which are *causally possible*. I could now weaken that to: . . . whether Q is true at the T-closest P-worlds which are *as nomologically similar to the actual world as any P-world is*. Where P itself is causally possible, the two versions are equivalent; but for counterlegal P the original theory is useless whereas the weakened one offers some hope of sorting out true from false by looking at the causally impossible P-worlds which are most like the actual world in their nomological structure.

That provision for counterlegals can easily be grafted onto the theory I have presented. I would accept it, if I had a workable concept of nomological similarity; but I haven't, and I don't think that anyone else has either.

If that Lewis-like approach to counterlegals cannot be made to work, we might tackle them within a Goodman-like framework. That is what is done by Pollock, who seems to be the only philosopher to have worked in detail on counterlegals (pp. 56–57, 93–97). Pollock does not speak of nomological similarity: his entire theory of counterfactuals has more in common with Goodman's than with Lewis's, and all through it he speaks not of maximally similar worlds but rather of worlds which can be reached from the actual world by 'minimal change'. Applying this to counterlegals, he says that if P is causally impossible then (P < Q) is true if Q is true at all P-worlds which are reachable from the actual world by a minimal change in what laws obtain. In giving details, he speaks of 'making deletions' in 'the actual set of basic laws', taking a minimal change to be one which deletes as few of the set's members as possible. That sounds all right until we remember to ask: how can we count basic laws? Pollock's theory needs an objective way of individuating laws, determining what counts as one law rather than more than one (a disguised conjunction) or less than one (a disguised disjunction); and nothing in his book provides the means for doing this.[1] He faces up to the analogous problem for propositions which are not laws, presenting a device which is supposed to sift out the conjunctions and disjunctions from amongst them, enabling us to count the remainder; but that device could not conceivably help us to individuate laws. I suspect that that problem is insoluble.

If I am wrong, and either or both of those two approaches to counterlegals could be soundly based, then I could build one of them into the theory I have presented in this paper. I am in no more trouble with counterlegals than are any other theorists of counterfactuals. Still, we are all in trouble, and I want to suggest a way out of it.

Consider first how counterlogicals are handled. Most of us think that if P is absolutely impossible then P entails Q

[1] Unless we take wholly seriously a passing implication that the laws are *sentences* (see the definition of 'maximal P-consistent subset' on p. 57). That would align Pollock with the position I am going to advocate, but I don't think it is his considered position.

for every Q, and so (P < Q) for every Q. Yet a speaker can say things of the form 'If conjunction were not commutative, then Q would be the case', and be right for some values of Q and wrong for others, just so long as he is talking about the power structure of some *system* of logic—some set of rules and independent axioms—and saying that Q is a theorem in the system which results from the original one by deleting the commutativity axiom. Such conditionals can be rescued from triviality only by being made relative to some formulation of logical truth: they cannot be nontrivially evaluated in terms of possible worlds, since all their antecedents are false at every world.

I propose that we treat counterlegals in the same way. It is true that for them there are some available worlds, namely those which are possible but not causally possible (whereas counterlogicals would require worlds which are possible but not possible). But those worlds do not help unless we know how to select from amongst them; and it seems that we don't. For counterlegals, then, we must turn away from worlds and propositions and have recourse to sentences.

Before leaving counterlegals, I should say a little about a sort of conditional which Peter van Inwagen has called to my attention—a sort including 'If I reached Jupiter within the next ten seconds, that would be a miracle' and 'If Einstein's most famous statement were false, things would happen which in fact can't happen'. The problem does not concern selection from among the causally impossible worlds: in each case the consequent is so weak as to be true at every such world. What is troublesome about these conditionals is that they seem to be true and to have antecedents which are causally possible and consequents which are not! If that really is how they are, then they make trouble for any theory of counterfactuals, but most obviously and directly for theories which say, as mine does, that where P is possible (P < Q) must be evaluated purely in terms of causally possible worlds.

Fortunately, there is a plausible theory-saving manoeuvre, namely to say that when we accept such a conditional we are interpreting its antecedent as short-hand for something causally impossible. That is, we take the speaker to mean something like 'If I reached Jupiter within the next ten seconds from my present position millions of miles from Jupiter, then. . . ', and 'If Einstein's most famous statement, namely that $E = mc^2$, were false, then. . . '. I am supposing that we take these extra bits to be meant by the speaker; it won't do merely to take them to be true. If at T someone utters 'If I reached Jupiter within the next ten seconds, that would be a miracle' and this is to come out true, then the antecedent must not pick out all the worlds where the speaker reaches Jupiter at T + 10 seconds, but only those where the speaker is very far from Jupiter at T and he reaches Jupiter at T + 10 seconds. The closest members of the former class of worlds will be ones where at T the speaker is close to Jupiter at T, and no miracle occurs, thus making the conditional false. Thus, for the conditional to be true its antecedent must be stronger than it looks; that is, the speaker must mean a richer antecedent than he actually utters.

Something like that, I suggest, explains every conditional whose consequent is causally impossible and which seems to be true and to have a causally possible antecedent. The Einstein conditional might be thought to depend upon the speaker's making a *de re* reference to Einstein's dictum, but that is not really the point. The conditional is true just so long as the speaker includes in his meaning something that makes his antecedent causally impossible—that Einstein's most famous statement was $E = mc^2$, or merely that if was the expression of a law.

15. Do we need similarity?

If my theory is adopted, what becomes of the concept of similarity? That depends on how T-closeness is analysed. I have conceded that whatever T-closeness is, it had better imply that T-closest P-worlds will be similar to the actual world; and it may turn out that similarity must be used to define T-closeness. But of course one would prefer, if possible, to replace similarity by some more specific relations upon which similarity is supervenient. Rather than looking for a world which globally resembles the actual world, it would be better to look for one which shares with it every actually true proposition which meets condition ϕ. If only we could devise a ϕ which does the job! Goodman sought to find one, and confessed failure.

Actually, the two problems which brought him to a halt have now been solved. One was not really a problem to begin with: Goodman was worried about having to speak of causally possible worlds, or of laws, thinking that causal possibility must be elucidated with help from counterfactuals; but like most philosophers today I disagree with that, and shall say no more about it.[1] The other problem was one part of the problem about 'cotenability': it presented itself in Goodman's paper as the question of how to justify saying

> 'If this match had been struck, it would have lit', because the match was dry, the air was still, etc.

rather than saying

> 'If this match had been struck, it would have been wet', because the match did not light, the air was still, etc.

Since Goodman's paper appeared, there have been many solutions to this, or rather versions of a single solution. The core of it is to say that in evaluating (P < Q) you should start with the how the actual world is at the time (T) to which P pertains, proceeding from that to later times only by law-based inferences, not adding in facts taken from the actual subsequent course of events unless they are—in the manner discussed late in section 10 above—causally independent of the antecedent. That allows 'If the match had been struck (at T) it would have lit (at T + d)' on the grounds that the match was dry at T, but it does not allow 'If the match had been struck (at T) it would have been wet (at T)' on the grounds that the match did not light at T + d. This solution of Goodman's 'match' problem, vividly present in my theory, is also an ingredient in several others.

It is a limited solution, though, for all it does is to stop decisions about cotenability at T from being interfered with by facts pertaining to times other than T; for example, to stop the fact that the match did not light at T + d from intruding into the consideration of what truths about T are cotenable with 'The match was struck at T'. Other problems about cotenability are not helped by that confinement to the time of the antecedent, because they arise within those confines. The general problem about cotenability is this: given that P is false, we must delete some of the truth about the actual world at T to get a T-closest P-world; but do we just peel not-P off the surface or do we dig it out by the roots? Suppose that at T Jones is in neither of the Carolinas, and is utterly neutral as between them—he is not outlawed in North Carolina, or only a mile from the southern border of South Carolina, or anything like that. And suppose we want to evaluate conditionals of the form 'If Jones were in one of the Carolinas, then. . . '. What worlds should be looked at?

[1] Except to remark that anyone who thinks that the concept of law *could* be analyzed with help from counterfactuals, and who holds what used to be the popular view of how the analysis should go, is advised to consider the powerful objections raised in Peter van Inwagen's 'Laws and Counterfactuals', *Noûs* 13 (1979): 439–453.

They must not share with the actual world the truth that Jones is in neither of the Carolinas, obviously; but can they share with it the truth that Jones is not in North Carolina? If so, then we get 'If Jones were in one, he would be in South Carolina', and by parity of reasoning we also get 'If Jones were in one, he would be in North Carolina.' Goodman's only solution was to ban any conditional which has an equally good rival, as each of these has. But if that 'solution' is given its head, it will falsify almost every true counterfactual, as Goodman agreed when this was said and defended by W. T. Parry.[1] Here is how that argument goes.

Take any true conditional (P < Q) which owes its truth to the truth of a certain proposition R. Goodman will not let us use R to get from P to Q if there is another true proposition R′ which would take us from P to not-Q. And there always *is* such an R′: for many cases it is simply 'Either not-P or not-R.' For example, we may want to say that if the match had been struck it would have lit, because (R) *it was dry*; but there is the rival claim that if the match had been struck it would not have lit, because (R′) *either it was not struck or it was not dry*. Goodman's rule makes those two rivals kill one another off; so it denies truth to the perfectly acceptable conditional (struck < lit), and by similar reasoning it denies truth to virtually every counterfactual.

That example, in which Goodman's rule comes to grief, has just the same logical structure as his own Carolina example, which is supposed to show the rule to advantage. Faced with someone who asserts (Carolina < South Carolina) on the grounds of the true (R) 'Not North Carolina', Goodman rejects that because of the rival (Carolina < North Carolina) based on the equally true (R′) 'Not South Carolina'—and the latter is equivalent to 'Either not Carolina or North Carolina', which is 'Either not-P or not-R'.

Pollock's Goodmanian theory incorporates Goodman's solution to this 'Carolina problem', as we might call it, together with a device which is supposed to stop the solution from indulging in overkill. In deciding which worlds are the closest, Pollock says, we must look only at the 'simple' propositions which are true at the actual world, not at complex propositions; for example, negative and disjunctive ones. That keeps at bay the disqualifying rivals which were routinely available, since the construction of them depended on negation and disjunction. If I were content to work with Pollock's notion of 'simplicity', as expounded on his pages 91–93, I would have a complete analysis of counterfactual conditionals: it would handle T-closeness in the Goodman-Pollock manner, and go on from there in the manner described in this paper. But I don't find Pollock's 'simplicity' convincing. *Sentences* split into disjunctive and nondisjunctive, but do *propositions*?

Slote's Goodmanian theory tries in a different way to solve a variant on the Carolina problem (see his pp. 15–17)—the variation rules out a certain solution to the problem in its simple form, but I shan't go into that here. Slote's solution relies upon something he calls a 'despiteness relation' between propositions, and he defines this label in terms of the notion of 'a valid explanation (involving no extraneous elements)', a notion which is not analyzed in turn. I think it is fair to say that this makes Slote's analysis less deep and objective and illuminating than one would like it to be.

Jackson seems not to have known of the continuing existence of the Carolina problem. His Goodmanian account of how to locate the T-closest P-worlds contains no *prima facie* solution to it (see his p. 19).

[1] See Parry (1957), and Nelson Goodman, 'Parry on Counterfactuals', *The Journal of Philosophy* 54 (1957): pp. 442–445. I am indebted to David Sanford for calling these discussions to my attention.

On the other hand, a similarity-based theory can take the Carolina problem in its stride. If Jones is in neither of the Carolinas, and is not outlawed in one or devoted to one or the like, then worlds where he is in North Carolina are neither more nor less like the actual world than are ones where he is in South Carolina, and so neither unwanted conditional comes out as true: they do indeed kill one another off. But it is not in general the case that where P is false and R true, and it is reasonable to assume that at the T-closest P-world (P & R) is true, there will be an equally good case for supposing that the rival (P & (either not-P or not-R)) is true at those worlds; for it may be obvious that worlds where (P & R) is true are more like the actual world than ones where (P & (either not-P or not-R)) are true.

16. Similarity and worlds

It would be good to have the Carolina problem solved—not merely so as to get a complete analysis but also so as to know where we stand with regard to the concept of a possible world. I shall explain.

Any viable theory of counterfactuals can be expressed in the language of 'possible worlds.' For example, the clause '. . . Q is derivable from P by means of causal laws. . . ' can be expressed in the form '. . . Q is true at every causally possible P-world. . . ', and similarly for other elements in Goodmanian theories.

So nothing hangs on the difference between analyses which *do* and ones which *don't* mention possible worlds. What is significant is the line between ones which *must* mention them and ones which *needn't*. Although Goodmanian theories can be expressed in that way, they need not be, for they can instead be expressed in the form

> (P < Q) is true ≡ Q is derivable from (P & R) for some true R such that Φ(P, Q, R) ,

where all the outstanding problems are packed into Φ. There is no pressure on us to re-express that in terms of worlds. Compare that with an analysis of the form

> (P < Q) is true ≡ Q is true at the P-worlds which are most Θ-similar to the actual world,

where the outstanding problems are packed into Θ, and where the concept of similarity is being given a basic use, not merely brought in as supervenient on relations which are expressible in other ways. An adherent of such an analysis does need the concept of a possible world, for he rests basic weight on a similarity relation which needs worlds as relata. This fact, which was first pointed out by Jackson (pp. 18–19), is of great importance *if* it matters a lot to know whether counterfactuals can be analyzed without recourse to possible worlds. It certainly matters a bit: if worlds are needed, that explains why philosophers were defeated by counterfactuals in the middle years of the 20th century; to succeed they needed similarity, which required an ontology containing worlds, which did not re-enter philosophy until later.

But that is a merely historical point; and I don't know how much the question matters philosophically. Anyway, as far as this present paper goes, I am not committed either way. Since I prefer sharp edges and real understanding to smooth surfaces and mere truth, I would of course prefer to dispense with similarity, thus joining company with Goodman, Chisholm, Jackson, Pollock, Slote, and others. But we cannot always have as much understanding as we would like: what I would prefer may prove to be impossible, in which case we must explain T-closeness through similarity, in the manner of Lewis, Stalnaker (apparently), Davis, Bigelow, and others. That has no effect on my arguments in this paper. Whatever the truth about T-closeness turns out to be, I contend for a symmetrical, unified theory in which counterfactuals may run freely in either temporal direction,

not deterred by the Downing scare story or side-tracked by hopes of explaining time's arrow.[1]

Bibliography

Jonathan Bennett, 'Counterfactuals and Possible Worlds', *Canadian Journal of Philosophy* 4 (1974): pp. 381–402.

John Bigelow, 'If-Then Meets Possible Words', *Philosophia* 6 (1976): pp. 215–235.

Roderick M. Chisholm, 'The Contrary-to-Fact Conditional', *Mind* 55 (1946): pp. 289–307.

Wayne A. Davis, 'Indicative and Subjunctive Conditionals', *The Philosophical Review* 88 (1979): pp. 544–564.

P. B. Downing, 'Subjunctive Conditionals, Time Order, and Causation', *Proceedings of the Aristotelian Society* 59 (1958–59): pp. 125–140.

Kit Fine, Review of Lewis's *Counterfactuals*, *Mind* 84 (1975): pp. 451–458.

Nelson Goodman, 'The Problem of Counterfactual Conditionals', *Journal of Philosophy* 44 (1947): pp. 113–128.

Frank Jackson, 'A Causal Theory of Counterfactuals', *Australasian Journal of Philosophy* 55 (1977): pp. 3–21.

David Lewis, *Counterfactuals* (Princeton, 1973).

David Lewis, 'Counterfactual Dependence and Time's Arrow', *Nôus* 13 (1979): pp. 455–476.

William Tuthill Parry, 'Reexamination of the Problem of Counterfactual Conditionals', *The Journal of Philosophy* 54 (1957): pp. 85–94.

John Pollock, *Subjunctive Reasoning* (Dordrecht, 1976).

Michael A. Slote, 'Time in Counterfactuals', *The Philosophical Review* 87 (1978): pp. 3–27.

Robert C. Stalnaker, 'A Theory of Conditionals', *American Philosophical Quarterly*, monograph no. 2 (1968): pp. 98–112.

[1] In writing this paper I have been much helped by the comments of David Lewis, Frank Jackson, Peter van Inwagen, Thomas McKay, Richmond Thomason, Sylvain Bromberger, David Sanford, and the editors of *The Philosophical Review*; by a series of conversations with Mark Heller; and by discussions at M.I.T., Brandeis University, Wayne State University, the University of Colorado at Boulder, the University of California at Berkeley, and a colloquium at the University of North Carolina. I am indebted to Carl Matheson for making me aware that I may be on a collision course with some popular beliefs about individual essences—beliefs about 'the necessity of origins'—but that is a topic for a further paper.