Counterfactual Dependence and Time's Arrow

DAVID LEWIS

PRINCETON UNIVERSITY

THE ASYMMETRY OF COUNTERFACTUAL DEPENDENCE

Today I am typing words on a page. Suppose today were different. Suppose I were typing different words. Then plainly tomorrow would be different also; for instance, different words would appear on the page. Would yesterday also be different? If so, how? Invited to answer, you will perhaps come up with something. But I do not think there is anything you can say about how yesterday would be that will seem clearly and uncontroversially true.

The way the future is depends counterfactually on the way the present is. If the present were different, the future would be different; and there are counterfactual conditionals, many of them as unquestionably true as counterfactuals ever get, that tell us a good deal about how the future would be different if the present were different in various ways. Likewise the present depends counterfactually on the past, and in general the way things are later depends on the way things were earlier.

Not so in reverse. Seldom, if ever, can we find a clearly true counterfactual about how the past would be different if the present were somehow different. Such a counterfactual, unless clearly false, normally is not clear one way or the other. It is at best doubtful whether the past depends counterfactually on the present, whether the present depends on the future, and in general whether the way things are earlier depends on the way things will be later.

Often, indeed, we seem to reason in a way that takes it for granted that the past is counterfactually independent of the present: that is, that even if the present were different, the

past would be just as it actually is. In reasoning from a counterfactual supposition, we use auxiliary premises drawn from (what we take to be) our factual knowledge. But not just anything we know may be used, since some truths would not be true under the given supposition. If the supposition concerns the present, we do not feel free to use all we know about the future. If the supposition were true, the future would be different and some things we know about the actual future might not hold in this different counterfactual future. But we do feel free, ordinarily, to use whatever we know about the past. We evidently assume that even if our supposition about the present were true, the past would be no different. If I were acting otherwise just now, I would revenge a wrong done me last year—it is absurd even to raise the question whether that past wrong would have taken place if I were acting otherwise now! More generally, in reasoning from a counterfactual supposition about any time, we ordinarily assume that facts about earlier times are counterfactually independent of the supposition and so may freely be used as auxiliary premises.

I would like to present a neat contrast between counterfactual dependence in one direction of time and counterfactual independence in the other direction. But until a distinction is made, the situation is not as neat as that. There are some special contexts that complicate matters. We know that present conditions have their past causes. We can persuade ourselves, and sometimes do, that if the present were different then these past causes would have to be different, else they would have caused the present to be as it actually is. Given such an argument—call it a *back-tracking argument*—we willingly grant that if the present were different, the past would be different too. I borrow an example from Downing ([5]). Jim and Jack quarreled yesterday, and Jack is still hopping mad. We conclude that if Jim asked Jack for help today, Jack would not help him. But wait: Jim is a prideful fellow. He never would ask for help after such a quarrel; if Jim were to ask Jack for help today, there would have to have been no quarrel yesterday. In that case Jack would be his usual generous self. So if Jim asked Jack for help today, Jack would help him after all.

At this stage we may be persuaded (and rightly so, I think) that if Jim asked Jack for help today, there would have been no quarrel yesterday. But the persuasion does not last. We

very easily slip back into our usual sort of counterfactual reasoning, and implicitly assume once again that facts about earlier times are counterfactually independent of facts about later times. Consider whether pride is costly. In this case, at least, it costs Jim nothing. It would be useless for Jim to ask Jack for help, since Jack would not help him. We rely once more on the premise we recently doubted: if Jim asked Jack for help today, the quarrel would nevertheless have taken place yesterday.

What is going on, I suggest, can best be explained as follows. (1) Counterfactuals are infected with vagueness, as everyone agrees. Different ways of (partly) resolving the vagueness are appropriate in different contexts. Remember the case of Caesar in Korea: had he been in command, would he have used the atom bomb? Or would he have used catapults? It is right to say either, though not to say both together. Each is true under a resolution of vagueness appropriate to some contexts. (2) We ordinarily resolve the vagueness of counterfactuals in such a way that counterfactual dependence is asymmetric (except perhaps in cases of time travel or the like). Under this standard resolution, back-tracking arguments are mistaken: if the present were different the past would be the same, but the same past causes would fail somehow to cause the same present effects. If Jim asked Jack for help today, somehow Jim would have overcome his pride and asked despite yesterday's quarrel. (3) Some special contexts favor a different resolution of vagueness, one under which the past depends counterfactually on the present and some back-tracking arguments are correct. If someone propounds a back-tracking argument, for instance, his cooperative partners in conversation will switch to a resolution that gives him a chance to be right. (This sort of accommodating shift in abstract features of context is common; see Lewis ([14]).) But when the need for a special resolution of vagueness comes to an end, the standard resolution returns. (4) A counterfactual saying that the past would be different if the present were somehow different may come out true under the special resolution of its vagueness, but false under the standard resolution. If so, call it a *back-tracking counterfactual*. Taken out of context, it will not be clearly true or clearly false. Although we tend to favor the standard resolution, we also charitably tend to favor a resolution which gives the sentence under consideration a chance of truth.

(Back-tracking counterfactuals, used in a context that favors their truth, are marked by a syntactic peculiarity. They are the ones in which the usual subjunctive conditional constructions are readily replaced by more complicated constructions: "If it were that. . . then it would have to be that. . ." or the like. A suitable context may make it acceptable to say "If Jim asked Jack for help today, there would have been no quarrel yesterday", but it would be more natural to say ". . . there would have to have been no quarrel yesterday." Three paragraphs ago, I used such constructions to lure you into a context that favors back-tracking.)

I have distinguished the standard resolution of vagueness from the sort that permits back-tracking only so that I can ask you to ignore the latter. Only under the standard resolution do we have the clear-cut asymmetry of counterfactual dependence that interests me.

I do not claim that the asymmetry holds in all possible, or even all actual, cases. It holds for the sorts of familiar cases that arise in everyday life. But it well might break down in the different conditions that might obtain in a time machine, or at the edge of a black hole, or before the Big Bang, or after the Heat Death, or at a possible world consisting of one solitary atom in the void. It may also break down with respect to the immediate past. We shall return to these matters later.

Subject to these needed qualifications, what I claim is as follows. Consider those counterfactuals of the form "If it were that A, then it would be that C" in which the supposition A is indeed false, and in which A and C are entirely about the states of affairs at two times t_A and t_C respectively. Many such counterfactuals are true in which C also is false, and in which t_C is later than t_A. These are counterfactuals that say how the way things are later depends on the way things were earlier. But if t_C is earlier than t_A, then such counterfactuals are true if and only if C is true. These are the counterfactuals that tell us how the way things are earlier does not depend on the way things will be later.

ASYMMETRIES OF CAUSATION AND OPENNESS

The asymmetry of counterfactual dependence has been little discussed. (However, see Downing [5], Bennett [2], and Slote [19].) Some of its consequences are better known. It is instruc-

tive to see how the asymmetry of counterfactual dependence serves to explain these more familiar asymmetries.

Consider the temporal asymmetry of causation. Effects do not precede their causes, or at least not ordinarily. Elsewhere ([12]) I have advocated a counterfactual analysis of causation: (1) the relation of cause to effect consists in their being linked by a causal chain; (2) a causal chain is a certain kind of chain of counterfactual dependences; and (3) the counterfactuals involved are to be taken under the standard resolution of vagueness. If anything of the sort is right, there can be no backward causation without counterfactual dependence of past on future. Only where the asymmetry of counterfactual dependence breaks down can there possibly be exceptions to the predominant futureward direction of causation.

Consider also what I shall call the *asymmetry of openness*: the obscure contrast we draw between the "open future" and the "fixed past." We tend to regard the future as a multitude of alternative possibilities, a "garden of forking paths" in Borges' phrase, whereas we regard the past as a unique, settled, immutable actuality. These descriptions scarcely wear their meaning on their sleeves, yet do seem to capture some genuine and important difference between past and future. What can it be? Several hypotheses do not seem quite satisfactory.

Hypothesis 1: Asymmetry of Epistemic Possibility. Is it just that we know more about the past than about the future, so that the future is richer in epistemic possibilities? I think that's not it. The epistemic contrast is a matter of degree, not a difference in kind, and sometimes is not very pronounced. There is a great deal we know about the future, and a great deal we don't know about the past. Ignorance of history has not the least tendency to make (most of) us think of the past as somewhat future-like, multiple, open, or unfixed.

Hypothesis 2: Asymmetry of Multiple Actuality. Is it that all our possible futures are equally actual? It is possible, I think, to make sense of multiple actuality. Elsewhere I have argued for two theses (in [9] and [8]): (1) any inhabitant of any possible world may truly call his own world actual; (2) we ourselves inhabit this one world only, and are not identical with our other-wordly counterparts. Both theses are controversial, so

perhaps I am right about one and wrong about the other. If (1) is true and (2) is false, here we are inhabiting several worlds at once and truly calling all of them actual. (Adams argues contrapositively in [1], arguing from the denial of multiple actuality and the denial of (2) to the denial of (1).) That makes sense, I think, but it gives us no asymmetry. For in some sufficiently broad sense of possibility, we have alternative possible pasts as well as alternative possible futures. But if (1) is true and (2) is false, that means that *all* our possibilities are equally actual, past as well as future.

Hypothesis 3: Asymmetry of Indeterminism. Is it that we think of our world as governed by indeterministic laws of nature, so that the actual past and present are nomically compossible with various alternative future continuations? I think this hypothesis also fails.

For one thing, it is less certain that our world is indeterministic than that there is an asymmetry between open future and fixed past—whatever that may turn out to be. Our best reason to believe in indeterminism is the success of quantum mechanics, but that reason is none too good until quantum mechanics succeeds in giving a satisfactory account of processes of measurement.

For another thing, such reason as we have to believe in indeterminism is reason to believe that the laws of nature are indeterministic in both directions, so that the actual future and present are nomically compossible with various alternative pasts. If there is a process of reduction of the wave packet in which a given superposition may be followed by any of many eigenstates, equally this is a process in which a given eigenstate may have been preceded by any of many superpositions. Again we have no asymmetry.

I believe that indeterminism is neither necessary nor sufficient for the asymmetries I am discussing. Therefore I shall ignore the possibility of indeterminism in the rest of this paper, and see how the asymmetries might arise even under strict determinism. A *deterministic* system of laws is one such that, whenever two possible worlds both obey the laws perfectly, then either they are exactly alike throughout all of time, or else they are not exactly alike through any stretch of time. They are alike always or never. They do not diverge, matching perfectly in their initial segments but not thereafter; neither

do they converge. Let us assume, for the sake of the argument, that the laws of nature of our actual world are in this sense deterministic.

(My definition of determinism derives from Montague ([15]), but with modifications. (1) I prefer to avoid his use of mathematical constructions as *ersatz* possible worlds. But should you prefer *ersatz* worlds to the real thing, that will not matter for the purposes of this paper. (2) I take exact likeness of worlds at times as a primitive relation; Montague instead uses the relation of having the same complete description in a certain language, which he leaves unspecified.

My definition presupposes that we can identify stretches of time from one world to another. That presupposition is questionable, but it could be avoided at the cost of some complication.)

Hypothesis 4: Asymmetry of Mutability. Is it that we can change the future, but not the past? Not so, if "change" has its literal meaning. It is true enough that if t is any past time, then we cannot bring about a difference between the state of affairs at t at time t_1 and the (supposedly changed) state of affairs at t at a later time t_2. But the pastness of t is irrelevant; the same would be true if t were present or future. Past, present, and future are alike immutable. What explains the impossibility is that such phrases as "the state of affairs at t at t_1" or "the state of affairs at t at t_2", if they mean anything, just mean: the state of affairs at t. Of course we cannot bring about a difference between that and itself.

Final Hypothesis: Asymmetry of Counterfactual Dependence. Our fourth hypothesis was closer to the truth than the others. What we *can* do by way of "changing the future" (so to speak) is to 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. That is something like change. We make a difference. But it is not literally change, since the difference we make is between actuality and other possibilities, not between successive actualities. The literal truth is just that the future depends counterfactually on the present. It depends, partly, on what we do now.

Likewise, something we ordinarily *cannot* do by way of "changing the past" 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 acted differently in the present. The past would be the same, however we acted now. The past does not at all depend on what we do now. It is counterfactually independent of the present.

In short, I suggest that the mysterious asymmetry between open future and fixed past is nothing else than the asymmetry of counterfactual dependence. The forking paths into the future—the actual one and all the rest—are the many alternative futures that would come about under various counterfactual suppositions about the present. The one actual, fixed past is the one past that would remain actual under this same range of suppositions.

TWO ANALYSES OF COUNTERFACTUALS

I hope I have now convinced you that an asymmetry of counterfactual dependence exists; that it has important consequences; and therefore that it had better be explained by any satisfactory semantic analysis of counterfactual conditionals. In the rest of this paper, I shall consider how that explanation ought to work.

It might work by fiat. It is an easy matter to build the asymmetry into an analysis of counterfactuals, for instance as follows.

Analysis 1. Consider a counterfactual "If it were that A, then it would be that C" where A is entirely about affairs in a stretch of time t_A. Consider all those possible worlds w such that:

(1) A is true at w;

(2) w is exactly like our actual world at all times before a transition period beginning shortly before t_A;

(3) w conforms to the actual laws of nature at all times after t_A; and

(4) during t_A and the preceding transition period, w differs no more from our actual world than it must to permit A to hold.

The counterfactual is true if and only if C holds at every such world w.

In short, take the counterfactual present (if t_A is now), avoiding gratuitous difference from the actual present; graft it smoothly onto the actual past; let the situation evolve according to the actual laws; and see what happens. An analysis close to Analysis 1 has been put forward by Jackson ([7]). Bennett ([2]), Bowie ([3]), and Weiner ([21]) have considered, but not endorsed, similar treatments.

Analysis 1 guarantees the asymmetry of counterfactual dependence, with an exception for the immediate past. Let C be entirely about a stretch of time t_C. If t_C is later than t_A, then C may very well be false at our world, yet true at the worlds that meet the conditions listed in Analysis 1. We have the counterfactuals whereby the affairs of later times depend on those of earlier times. But if t_C is before t_A, and also before the transition period, then C holds at worlds that meet condition (2) if and only if C is true at our actual world. Since C is entirely about something that does not differ at all from one of these worlds to another, its truth value cannot vary. Therefore, except for cases in which t_C falls in the transition period, we have the counterfactuals whereby the affairs of earlier times are independent of those of later times.

We need the transition period, and should resist any temptation to replace (2) by the simpler and stronger

(2*) w is exactly like our actual world at all times before t_A.

(2*) makes for abrupt discontinuities. Right up to t, the match was stationary and a foot away from the striking surface. If it had been struck at t, would it have travelled a foot in no time at all? No; we should sacrifice the independence of the immediate past to provide an orderly transition from actual past to counterfactual present and future. 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 intervals of time in cases that are not at all extraordinary.

Analysis 1 seems to fit a wide range of counterfactuals; and it explains the asymmetry of counterfactuals dependence,

though with one rather plausible exception. Should we be content? I fear not, for two reasons.

First, Analysis 1 is built for a special case. We need a supposition about a particular time, and we need a counterfactual taken under the standard resolution of vagueness. What shall we do with suppositions such as

> If kangaroos had no tails. . .
>
> If gravity went by the inverse cube of distance. . .
>
> If Collett had ever designed a Pacific. . .

which are not about particular times? Analysis 1 cannot cope as it stands, nor is there any obvious way to generalize it. At most we could give separate treatments of other cases, drawing on the cases handled by Analysis 1. (Jackson ([7]) does this to some extent.) Analysis 1 is not much of a start toward a uniform treatment of counterfactuals in general.

Second, Analysis 1 gives us more of an asymmetry than we ought to want. No matter how special the circumstances of the case may be, no provision whatever is made for actual or possible exceptions to the asymmetry (except in the transition period). That is too inflexible. Careful readers have thought they could make sense of stories of time travel (see my [13] for further discussion); hard-headed psychical researchers have believed in precognition; speculative physicists have given serious consideration to tachyons, advanced potentials, and cosmological models with closed timelike curves. Most or all of these phenomena would involve special exceptions to the normal asymmetry of counterfactual dependence. It will not do to declare them impossible *a priori*.

The asymmetry-by-fiat strategy of Analysis 1 is an instructive error, not a dead loss. Often we do have the right sort of supposition, the standard resolution of vagueness, and no extraordinary circumstances. Then Analysis 1 works as well as we could ask. The right analysis of counterfactuals needs to be both more general and more flexible. But also it needs to agree with Analysis 1 over the wide range of cases for which Analysis 1 succeeds.

The right general analysis of counterfactuals, in my opinion, is one based on comparative similarity of possible worlds. Roughly, a counterfactual is true if every world that makes the

antecedent true without gratuitous departure from actuality is a world that also makes the consequent true. Such an analysis is given in my [10] and [11]; here is one formulation.

> *Analysis 2.* A counterfactual "If it were that A, then it would be that C" is (non-vacuously) true if and only if some (accessible) world where both A and C are true is more similar to our actual world, overall, than is any world where A is true but C is false.

This analysis is fully general: A can be a supposition of any sort. It is also extremely vague. Overall similarity among worlds is some sort of resultant of similarities and differences of many different kinds, and I have not said what system of weights or priorities should be used to squeeze these down into a single relation of overall similarity. I count that a virtue. Counterfactuals are both vague and various. Different resolutions of the vagueness of overall similarity are appropriate in different contexts.

Analysis 2 (plus some simple observations about the formal character of comparative similarity) is about all that can be said in full generality about counterfactuals. While not devoid of testable content—it settles some questions of logic—it does little to predict the truth values of particular counterfactuals in particular contexts. The rest of the study of counterfactuals is not fully general. Analysis 2 is only a skeleton. It must be fleshed out with an account of the appropriate similarity relation, and this will differ from context to context. Our present task is to see what sort of similarity relation can be combined with Analysis 2 to yield what I have called the standard resolution of vagueness: one that invalidates back-tracking arguments, one that yields an asymmetry of counterfactual dependence except perhaps under special circumstances, one that agrees with Analysis 1, our asymmetry-by-fiat analysis, whenever it ought to.

But first, a word of warning! Do not assume that just any respect of similarity you can think of must enter into the balance of overall similarity with positive weight. The point is obvious for some respects of similarity, if such they be. It contributes nothing to the similarity of two gemstones that both are grue. (To be *grue* is to be green and first examined before 2000 A.D. or blue and not first examined before 2000

A.D.) But even some similarities in less gruesome respects may count for nothing. They may have zero weight, at least under some reasonable resolutions of vagueness. To what extent are the philosophical writings of Wittgenstein similar, overall, to those of Heidegger? I don't know. But here is one respect of comparison that does not enter into it at all, not even with negligible weight: the ratio of vowels to consonants.

(Bowie ([3]) has argued that if some respects of comparison counted for nothing, my assumption of "centering" in [10] and [11] would be violated: worlds differing from ours only in the respects that don't count would be as similar to our world as our world is to itself. I reply that there may not be any worlds that differ from ours only in the respects that don't count, even if there are some respects that don't count. Respects of comparison may not be entirely separable. If the writings of two philosophers were alike in every respect that mattered, they would be word-for-word the same; then they would have the same ratio of vowels to consonants.)

And next, another word of warning! It is all too easy to make offhand similarity judgments and then assume that they will do for all purposes. But if we respect the extreme shiftiness and context-dependence of similarity, we will not set much store by offhand judgments. We will be prepared to distinguish between the similarity relations that guide our offhand explicit judgments and those that govern our counterfactuals in various contexts.

Indeed, unless we are prepared so to distinguish, Analysis 2 faces immediate refutation. Sometimes a pair of counterfactuals of the following form seem true: "If A, the world would be very different; but if A and B, the world would not be very different." Only if the similarity relation governing counterfactuals disagrees with that governing explicit judgments of what is "very different" can such a pair be true under Analysis 2. (I owe this argument to Pavel Tichý and, in a slightly different form, to Richard J. Hall.) It seems to me no surprise, given the instability even of explicit judgments of similarity, that two different comparative similarity relations should enter into the interpretation of a single sentence.

The thing to do is not to start by deciding, once and for all, what we think about similarity of worlds, so that we can afterwards use these decisions to test Analysis 2. What that would test would be the combination of Analysis 2 with a foolish denial of the shiftiness of similarity. Rather, we must

use what we know about the truth and falsity of counterfactuals to see if we can find some sort of similarity relation—not necessarily the first one that springs to mind—that combines with Analysis 2 to yield the proper truth conditions. It is this combination that can be tested against our knowledge of counterfactuals, not Analysis 2 by itself. In looking for a combination that will stand up to the test, we must use what we know about counterfactuals to find out about the appropriate similarity relation—not the other way around.

THE FUTURE SIMILARITY OBJECTION

Several people have raised what they take to be a serious objection against Analysis 2. (It was first brought to my attention by Michael Slote; it occurs, in various forms, in [2],[3], [4], [6], [7], [17], [18], and [19].) Kit Fine ([6]: 452) states it as follows.

> The counterfactual "If Nixon had pressed the button there would have been a nuclear holocaust" is true or can be imagined to be so. Now suppose that there never will be a nuclear holocaust. Then that counterfactual is, on Lewis's analysis, very likely false. For given any world in which antecedent and consequent are both true it will be easy to imagine a closer world in which the antecedent is true and the consequent false. For we need only imagine a change that prevents the holocaust but that does not require such a great divergence from reality.

The presence or absence of a nuclear holocaust surely does contribute with overwhelming weight to some prominent similarity relations. (For instance, to one that governs the explicit judgment of similarity in the consequent of "If Nixon had pressed the button, the world would be very different.") But the relation that governs the counterfactual may not be one of these. It may nevertheless be a relation of overall similarity—not because it is likely to guide our explicit judgments of similarity, but rather because it is a resultant, under some system of weights or priorities, of a multitude of relations of similarity in particular respects.

Let us take the supposition that Nixon pressed the button as implicitly referring to a particular time t—let it be the darkest moment of the final days. Consider w_0, a world that may or may not be ours. At w_0, Nixon does not press the button at t and no nuclear holocaust ever occurs. Let w_0 also be

a world with deterministic laws, since we have confined our attention here to counterfactual dependence under determinism. Let w_0 also be a world that fits our worst fantasies about the button: there is such a button, it is connected to a fully automatic command and control stystem, the wired-in war plan consists of one big salvo, everything is in faultless working order, there is no way for anyone to stop the attack, and so on. Then I agree that Fine's counterfactual is true at w_0: if Nixon had pressed the button, there would have been a nuclear holocaust.

There are all sorts of worlds where Nixon (or rather, a counterpart of Nixon) presses the button at t. We must consider which of these differ least, under the appropriate similarity relation, from w_0. Some are non-starters. Those where the payload of the rockets consists entirely of confetti depart gratuitously from w_0 by any reasonable standards. The more serious candidates fall into several classes.

One class is typified by the world w_1. Until shortly before t, w_1 is exactly like w_0. The two match perfectly in every detail of particular fact, however minute. Shortly before t, however, the spatio-temporal region of perfect match comes to an end as w_1 and w_0 begin to diverge. The deterministic laws of w_0 are violated at w_1 in some simple, localized, inconspicuous way. A tiny miracle takes place. Perhaps a few extra neurons fire in some corner of Nixon's brain. As a result of this, Nixon presses the button. With no further miracles events take their lawful course and the two worlds w_1 and w_0 go their separate ways. The holocaust takes place. From that point on, at least so far as the surface of this planet is concerned, the two worlds are not even approximately similar in matters of particular fact. In short, the worlds typified by w_1 are the worlds that meet the conditions listed in Analysis 1, our asymmetry-by-fiat analysis. What is the case throughout these worlds is just what we think would have been the case if Nixon had pressed the button (assuming that we are at w_0, and operating under the standard resolution of vagueness). Therefore the worlds typified by w_1 should turn out to be more similar to w_0, under the similarity relation we seek, than any of the other worlds where Nixon pressed the button.

(When I say that a miracle takes place at w_1, I mean that there is a violation of the laws of nature. But note that the violated laws are not laws of the same world where they are

violated. That is impossible; whatever else a law may be, it is at least an exceptionless regularity. I am using "miracle" to express a relation between different worlds. A miracle at w_1, relative to w_0, is a violation at w_1 of the laws of w_0, which are at best the almost-laws of w_1. The laws of w_1 itself, if such there be, do not enter into it.)

A second class of candidates is typified by w_2. This is a world completely free of miracles: the deterministic laws of w_0 are obeyed perfectly. However, w_2 differs from w_0 in that Nixon pressed the button. By definition of determinism, w_2 and w_0 are alike always or alike never, and they are not alike always. Therefore they are not exactly alike through any stretch of time. They differ even in the remote past. What is worse, there is no guarantee whatever that w_2 can be chosen so that the differences diminish and eventually become negligible in the more and more remote past. Indeed, it is hard to imagine how two deterministic worlds anything like ours could possibly remain just a little bit different for very long. There are altogether too many opportunities for little differences to give rise to bigger differences.

Certainly such worlds as w_2 should not turn out to be the most similar worlds to w_0 where Nixon pressed the button. That would lead to back-tracking unlimited. (And as Bennett observes in [2], it would make counterfactuals useless; we know far too little to figure out which of them are true under a resolution of vagueness that validates very much back-tracking.) The lesson we learn by comparing w_1 and w_2 is that under the similarity relation we seek, a lot of perfect match of particular fact is worth a little miracle.

A third class of candidates is typified by w_3. This world begins like w_1. Until shortly before t, w_3 is exactly like w_0. Then a tiny miracle takes place, permitting divergence. Nixon presses the button at t. But there is no holocaust, because soon after t a second tiny miracle takes place, just as simple and localized and inconspicuous as the first. The fatal signal vanishes on its way from the button to the rockets. Thereafter events at w_3 take their lawful course. At least for a while, worlds w_0 and w_3 remain very closely similar in matters of particular fact. But they are no longer exacly alike. The holocaust has been prevented, but Nixon's deed has left its mark on the world w_3. There are his fingerprints on the button. Nixon is still trembling, wondering what went wrong—or right. His gin bottle is depleted. The click of the button has

been preserved on tape. Light waves that flew out the window, bearing the image of Nixon's finger on the button, are still on their way into outer space. The wire is ever so slightly warmed where the signal current passed through it. And so on, and on, and on. The differences between w_3 and w_0 are many and varied, although no one of them amounts to much.

I should think that the close similarity between w_3 and w_0 could not last. Some of the little differences would give rise to bigger differences sooner or later. Maybe Nixon's memoirs are more sanctimonious at w_3 than at w_0. Consequently they have a different impact on the character of a few hundred out of the millions who read them. A few of these few hundred make different decisions at crucial moments of their lives—and we're off! But if you are not convinced that the differences need increase, no matter. My case will not depend on that.

If Analysis 2 is to succeed, such worlds as w_3 must not turn out to be the most similar worlds to w_0 where Nixon pressed the button. The lesson we learn by comparing w_1 and w_3 is that under the similarity relation we seek, close but approximate match of particular fact (especially if it is temporary) is not worth even a little miracle. Taking that and the previous lesson of w_2 together, we learn that perfect match of particular fact counts for much more than imperfect match, even if the imperfect match is good enough to give us similarity in respects that matter very much to us. I do not claim that this pre-eminence of perfect match is intuitively obvious. I do not claim that it is a feature of the similarity relations most likely to guide our explicit judgments. It is not; else the objection we are considering never would have been put forward. (See also the opinion survey reported by Bennett in [2].) But the pre-eminence of perfect match is a feature of some relations of overall similarity, and it must be a feature of any similarity relation that will meet our present needs.

A fourth class of candidates is typified by w_4. This world begins like w_1 and w_3. There is perfect match with w_0 until shortly before t, there is a tiny divergence miracle, the button is pressed. But there is a widespread and complicated and diverse second miracle after t. It not only prevents the holocaust but also removes all traces of Nixon's button-pressing. The cover-up job is miraculously perfect. Of course the fatal signal vanishes, just as at w_3, but there is much more. The

fingerprint vanishes, and the sweat returns to Nixon's fingertip. Nixon's nerves are soothed, his memories are falsified, and so he feels no need of the extra martini. The click on the tape is replaced by innocent noises. The receding light waves cease to bear their incriminating images. The wire cools down, and not by heating its surroundings in the ordinary way. And so on, and on, and on. Not only are there no traces that any human detective could read; in every detail of particular fact, however minute, it is just as if the button-pressing had never been. The worlds w_4 and w_0 reconverge. They are exactly alike again soon after t, and exactly alike forevermore. All it takes is enough of a reconvergence miracle: one involving enough different sorts of violations of the laws of w_0, in enough different places. Because there are many different sorts of traces to be removed, and because the traces spread out rapidly, the cover-up job divides into very many parts. Each part requires a miracle at least on a par with the small miracle required to prevent the holocaust, or the one required to get the button pressed in the first place. Different sorts of unlawful processes are needed to remove different sorts of traces: the miraculous vanishing of a pulse of current in a wire is not like the miraculous rearrangement of magnetized grains on a recording tape. The big miracle required for perfect reconvergence consists of a multitude of little miracles, spread out and diverse.

Such worlds as w_4 had better not turn out to be the most similar worlds to w_0 where Nixon pressed the button. The lesson we learn by comparing w_1 and w_4 is that under the similarity relation we seek, perfect match of particular fact even through the entire future is not worth a big, widespread, diverse miracle. Taking that and the lesson of w_2 together, we learn that avoidance of big miracles counts for much more than avoidance of little miracles. Miracles are not all equal. The all-or-nothing distinction between worlds that do and that do not ever violate the laws of w_0 is not sensitive enough to meet our needs.

This completes our survey of the leading candidates. There are other candidates, but they teach us nothing new. There are some worlds where approximate reconvergence to w_0 is secured by a second small miracle before t, rather than afterward as at w_3: Haig has seen fit to disconnect the button. Likewise there are worlds where a diverse and widespread miracle to permit perfect reconvergence takes place mostly

before and during t: Nixon's fingers leave no prints, the tape recorder malfunctions, and so on.

Under the similarity relation we seek, w_1 must count as closer to w_0 than any of w_2, w_3, and w_4. That means that a similarity relation that combines with Analysis 2 to give the correct truth conditions for counterfactuals such as the one we have considered, taken under the standard resolution of vagueness, must be governed by the following system of weights or priorities.

(1) It is of the first importance to avoid big, widespread, diverse violations of law.

(2) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.

(3) It is of the third importance to avoid even small, localized, simple violations of law.

(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

(It is a good question whether approximate similarities of particular fact should have little weight or none. Different cases come out differently, and I would like to know why. Tichý ([20]) and Jackson ([7]) give cases which appear to come out right under Analysis 2 only if approximate similarities count for nothing; but Morgenbesser has given a case, reported in Slote ([19]), which appears to go the other way. This problem was first brought my attention by Ernest Loevinsohn.)

Plenty of unresolved vagueness remains, of course, even after we have distinguished the four sorts of respect of comparison and ranked them in decreasing order of importance. But enough has been said to answer Fine's objection; and I think other versions of the future similarity objection may be answered in the same way.

THE ASYMMETRY OF MIRACLES

Enough has been said, also, to explain why there is an asymmetry of counterfactual dependence in such a case as we have

just considered. If Nixon had pressed the button, the future would have been of the sort found at w_1: a future very different, in matters of particular fact, from that of w_0. The past also would have been of the sort found at w_1: a past exactly like that of w_0 until shortly before t. Whence came this asymmetry? It is not built into Analysis 2. It is not built into the standards of similarity that we have seen fit to combine with Analysis 2.

It came instead from an asymmetry in the range of candidates. We considered worlds where a small miracle permitted divergence from w_0. We considered worlds where a small miracle permitted approximate convergence to w_0 and worlds where a big miracle permitted perfect convergence to w_0. But we did not consider any worlds where a small miracle permitted perfect convergence to w_0. If we had, our symmetric standards of similarity would have favored such worlds no less than w_1.

But are there any such worlds to consider? What could they be like: how could one small, localized, simple miracle possibly do all that needs doing? How could it deal with the fatal signal, the fingerprints, the memories, the tape, the light waves, and all the rest? I put it to you that it can't be done! Divergence from a world such as w_0 is easier than perfect convergence to it. Either takes a miracle, since w_0 is deterministic, but convergence takes very much more of a miracle. The asymmetry of counterfactual dependence arises because the appropriate standards of similarity, themselves symmetric, respond to this asymmetry of miracles.

It might be otherwise if w_0 were a different sort of world. I do not mean to suggest that the asymmetry of divergence and convergence miracles holds necessarily or universally. For instance, consider a simple world inhabited by just one atom. Consider the worlds that differ from it in a certain way at a certain time. You will doubtless conclude that convergence to this world takes no more of a varied and widespread miracle than divergence from it. That means, if I am right, that no asymmetry of counterfactual dependence prevails at this world. Asymmetry-by-fiat analyses go wrong for such simple worlds. The asymmetry of miracles, and hence of counterfactual dependence, rests on a feature of worlds like w_0 which very simple worlds cannot share.

ASYMMETRY OF OVERDETERMINATION

Any particular fact about a deterministic world is predeter-

mined throughout the past and postdetermined throughout the future. At any time, past or future, it has at least one *determinant*: a minimal set of conditions jointly sufficient, given the laws of nature, for the fact in question. (Members of such a set may be causes of the fact, or traces of it, or neither.) The fact may have only one determinant at a given time, disregarding inessential differences in a way I shall not try to make precise. Or it may have two or more essentially different determinants at a given time, each sufficient by itself. If so, it is *overdetermined* at that time. Overdetermination is a matter of degree: there might be two determinants, or there might be very many more than two.

I suggest that what makes convergence take so much more of a miracle than divergence, in the case of a world such as w_0, is an asymmetry of overdetermination at such a world. How much overdetermination of later affairs by earlier ones is there at our world, or at a deterministic world which might be ours for all we know? We have our stock examples—the victim whose heart is simultaneously pierced by two bullets, and the like. But those cases seem uncommon. Moreover, the overdetermination is not very extreme. We have more than one determinant, but still not a very great number. Extreme overdetermination of earlier affairs by later ones, on the other hand, may well be more or less universal at a world like ours. Whatever goes on leaves widespread and varied traces at future times. Most of these traces are so minute or so dispersed or so complicated that no human detective could ever read them; but no matter, so long as they exist. It is 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. (Even if a trace could somehow have been faked, traces of the absence of the requisite means of fakery may be included with the trace itself to form a set jointly sufficient for the fact in question.) If so, the abundance of future traces makes for a like abundance of future determinants. We may reasonably expect overdetermination toward the past on an altogether different scale from the occasional case of mild overdetermination toward the future.

That would explain the asymmetry of miracles. It takes a miracle to break the link between any determinant and that which it determines. Consider our example. To diverge from

w_0, a world where Nixon presses the button need only break the links whereby certain past conditions determine that he does not press it. To converge to w_0, a world where Nixon presses the button must break the links whereby a varied multitude of future conditions vastly overdetermine that he does not press it. The more overdetermination, the more links need breaking and the more widespread and diverse must a miracle be if it is to break them all.

An asymmetry noted by Popper ([16]) is a special case of the asymmetry of overdetermination. There are processes in which a spherical wave expands outward from a point source to infinity. The opposite processes, in which a spherical wave contracts inward from infinity and is absorbed, would obey the laws of nature equally well. But they never occur. A process of either sort exhibits extreme overdetermination in one direction. Countless tiny samples of the wave each determine what happens at the space-time point where the wave is emitted or absorbed. The processes that occur are the ones in which this extreme overdetermination goes toward the past, not those in which it goes toward the future. I suggest that the same is true more generally.

Let me emphasize, once more, that the asymmetry of overdetermination is a contingent, *de facto* matter. Moreover, it may be a local matter, holding near here but not in remote parts of time and space. If so, then all that rests on it—the asymmetries of miracles, of counterfactual dependence, of causation and openness—may likewise be local and subject to exceptions.

I regret that I do not know how to connect the several asymmetries I have discussed and the famous asymmetry of entropy.[1]

REFERENCES

[1] Robert M. Adams, "Theories of Actuality," NOÛS 8(1974): 211-31.
[2] Jonathan Bennett, review of Lewis ([10]), *The Canadian Journal of Philosophy* 4(1974): 381-402.
[3] G. Lee Bowie, "The Similarity Approach to Counterfactuals: Some Problems," NOÛS 13(1979): 484.
[4] Lewis Creary and Christopher Hill, review of Lewis ([10]), *Philosophy of Science* 42(1975): 341-4.
[5] P. B. Downing, "Subjunctive Conditionals, Time Order, and Causation," *Proceedings of the Aristotelian Society* 59(1959): 125-40.
[6] Kit Fine, review of Lewis ([10]), *Mind* 84(1975): 451-8.
[7] Frank Jackson, "A Causal Theory of Counterfactuals," *Australasian Journal of Philosophy* 55(1977): 3-21.

[8] David Lewis, "Counterpart Theory and Quantified Modal Logic," *Journal of Philosophy* 65(1968): 113-26.
[9] ———, "Anselm and Actuality," NOÛS 4(1970): 175-88.
[10] ———, *Counterfactuals* (Oxford: Blackwell, 1973).
[11] ———, "Counterfactuals and Comparative Possibility," *Journal of Philosophical Logic* 2(1973): 418-46.
[12] ———, "Causation," *Journal of Philosophy* 70(1973): 556-67; reprinted in Ernest Sosa (ed.), *Causation and Conditionals* (London: Oxford University Press, 1975).
[13] ———, "The Paradoxes of Time Travel," *American Philosophical Quarterly* 13(1976): 145-52.
[14] ———, "Scorekeeping in a Language Game," *Journal of Philosophical Logic*, (forthcoming).
[15] Richard Montague, "Deterministic Theories," in *Decisions, Values and Groups* (Oxford: Pergamon Press, 1962); reprinted in Montague, *Formal Philosophy* (New Haven: Yale University Press, 1974).
[16] Karl Popper, "The Arrow of Time," *Nature* 177(1956): 538.
[17] Tom Richards, "The Worlds of David Lewis," *Australasian Journal of Philosophy* 53(1975): 105-118.
[18] Eugene Schlossberger, "Similarity and Counterfactuals," *Analysis* 38(1978): 80-2.
[19] Michael A. Slote, "Time in Counterfactuals," *Philosophical Review* 87(1978): 3-27.
[20] Pavel Tichý, "A Counterexample to the Stalnaker-Lewis Analysis of Counterfactuals," *Philosophical Studies* 29(1976): 271-73.
[21] Joan Weiner, "Counterfactual Conundrum," NOÛS 13 (1979): 499-509.

NOTES

[1]I am grateful to many friends for discussion of these matters, and especially to Jonathan Bennett, Robert Goble, Philip Kitcher, Ernest Loevinsohn, John Perry, Michael Slote, and Robert Stalnaker. I am grateful to seminar audiences at several universities in New Zealand for comments on an early version of this paper, and to the New Zealand-United States Educational Foundation for making those seminars possible. I also thank Princeton University and the American Council of Learned Societies for research support at earlier stages. An earlier version of this paper was presented at the 1976 Annual Conference of the Australasian Association of Philosophy.