**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

# **Seventeen** Counterfactual Dependence and Time's Arrow

**David Lewis**

## 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

end p.32

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.

end p.33

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

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

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 co-operative 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

end p.34

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 instructive 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

end p.35

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

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

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-worldly 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 contra-positively 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.

end p.36

*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.

end p.37

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.

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

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.

end p.38

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

end p.39

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 counterfactual 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 . . .

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

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

end p.40

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

end p.41

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.

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

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

end p.42

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.

end p.43

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 system, 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

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

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;

end p.44

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 exactly alike. The holocaust has been prevented, but Nixon's deed has left its mark on the world $w_3$. There are his finger-prints 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

end p.45

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 wide-spread and complicated and diverse second miracle after $t$. It

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

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

end p.46

$w_4$ 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

end p.47

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 to 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

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

$w_0$. 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:

end p.48

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 necessary or universally. For instance, consider a simple word 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 predetermined 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

end p.49

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

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

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

end p.50

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): 477–98.

[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. Link

[8] David Lewis, "Counterpart Theory and Quantified Modal Logic," *Journal of Philosophy* 65(1968): 113–26. Link

[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; Link 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.

end p.51

[14] ——, "Scorekeeping in a Language Game," *Journal of Philosophical Logic* , 8 (1979): 339–59.

[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. Link

[17] Tom Richards, "The Worlds of David Lewis," *Australasian Journal of Philosophy* 53(1975): 105–118. Link

[18] Eugene Schlossberger, "Similarity and Counterfactuals," *Analysis* 38(1978): 80–2. Link

[19] Michael A. Slote, "Time in Counterfactuals," *Philosophical Review* 87(1978): 3–27. Link

[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.

**Postscripts to " 'Counterfactual Dependence and Time's Arrow"**

## A. New Theory and Old?

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

From time to time I have been told, much to my surprise, that this paper presents a "new theory" of counterfactuals, opposed to the "old theory" I had advanced in earlier writings.[1]

I would have thought, rather, that the truth of the matter was as follows. In the earlier writings I said that counterfactuals were governed in their truth conditions by comparative overall similarity of worlds, but that there was no one precisely fixed relation of similarity that governed all counterfactuals always. To the contrary, the governing similarity relation was both vague and context-dependent. Different contexts would select different ranges of similarity relations, probably without ever reaching full determinacy. In this paper I reiterate all that.

end p.52

Then I focus attention on some contexts in particular, and on the range of similarity relations that apply in such contexts. Thereby I add to my earlier discussion, but do not at all subtract from it. Yet not a few readers think I have taken something back. Why?

The trouble seems to be that a comparative relation of the sort that I now put forward—one that turns to some extent on the size of regions of perfect match, and to some extent on the scarcity in one world of events that violate the laws of another—is not at all what my earlier writings led these readers to expect. But why not? I think the trouble has three sources.

One source, I think, is entrenched doubt about the very idea of similarity. It is widely thought that *every* shared property, in the most inclusive possible sense of that word, is *prima facie* a respect of similarity: that things can be similar in respect of satisfying the same miscellaneously disjunctive formula, or in respect of belonging to the same utterly miscellaneous class. If so, then there's little to be said about comparative similarity. Any two things, be they two peas in a pod or be they a raven and a writing-desk, are alike in infinitely many respects and unlike in equally many.

Against this scepticism, I observed that we undeniably do make judgments of comparative overall similarity. And readers took the point—but in far too limited a way. "Yes," I think they thought, "there is indeed a comparative relation that is special in the way it governs our explicit snap judgments. We can scarcely doubt that—we have an operational test. But leave that firm ground, and we're as much at sea as ever. Apart from that one special case, we do not understand how one shared property can be more or less of a similarity-maker than another; or how it can be that some orderings are comparisons of similarity and others aren't." And so I speak of similarity, and these sceptics understand me in the only way they can: they seize on the one discrimination they regard as unproblematic, since they can understand how to pick out one similarity relation operationally in terms of snap judgments. Then they observe, quite rightly, that the "similarity relation" I now put forward as governing counterfactuals isn't *that* one.

The right lesson would have been more far-reaching. Our ability to make the snap judgment is one reason, among others, to reject the sceptical, egalitarian orthodoxy. It just isn't so that all properties (in the most inclusive sense) are equally respects of similarity. Then it is by no means empty to say as I do that a relation of overall similarity is any weighted resultant of respects of similarity and dissimilarity. (To

end p.53

which I add that the weighting might be nonarchimedean; that is, we might have a system of priorities rather than trade-offs.) Here we have a class of comparative relations that can go far beyond the one that governs the snap judgments; and that yet falls far short of the class delineated just by the formal character of comparative similarity.

Once we reject egalitarianism, what shall we put in its place? An analysis, somehow, of the difference between those properties that are respects of similarity and those that aren't? A primitive distinction? A distinction built into our ontology, in the form of a denial of the very existence of the alleged properties that aren't respects of similarity? A fair question; but one it is risky to take up, lest we put the onus on the wrong side. What we know best on this subject, I think, is that egalitarianism is *prima facie* incredible. We are entitled to reject it without owing any developed alternative.[2]

A second source of trouble, I suspect, is that some readers think of imperfect similarity always as imperfect match, and neglect the case of perfect match over a limited region. To illustrate, consider three locomotives: 2818, 4018, and 6018. 2818 and 4018 are alike in this way: they have duplicate boilers, smokeboxes, and fireboxes (to the extent that two of a kind from an early 20th century production line ever are duplicates), and various lesser fittings also are duplicated. But 2818 is a slow, small-wheeled, two-cylindered 2–8–0 coal hauler—plenty of pull, little speed—whereas 4018 is the opposite, a fast, large-wheeled, four-cylindered 4–6–0 express passenger locomotive. So is 6018; but 6018, unlike 2818, has few if any parts that duplicate the corresponding parts of 4018. (6018 is a scaled-up and modernized version of 4018.) Anyone can see the way in which 6018 is more similar to 4018 than 2818 is. But I would insist that there is another way of comparing similarity, equally deserving of that name, on which the duplicate standard parts make 2818 the stronger candidate.

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

A third source of trouble may be a hasty step from similarity with respect to laws of nature to similarity *of* the laws—or, I might even say, to similarity of the linguistic codifications of laws. Consider three worlds. The first has some nice, elegant system of uniform laws. The second does not: the best way to write down its laws would be to write

end p.54

down the laws of the first world, then to mutilate them by sticking in clauses to permit various exceptions in an unprincipled fashion. Yet almost everything that ever happens in the second world conforms perfectly to the laws of the first. The third world does have a nice, elegant, uniform system; its laws resemble those of the first world except for a change of sign here, a switch from inverse square to inverse cube there, and a few other such minor changes. Consequently, the third world constantly violates the laws of the first; any little thing that goes on in the third would be prohibited by the laws of the first. Focus on the linguistic codification of the laws, and it may well seem that the third world resembles the first with respect to laws far more than the second does. But I would insist that there is another way of comparing similarity with respect to laws, equally deserving of that name, on which the second world resembles the first very well, and the third resembles the first very badly. That is the way that neglects linguistic codifications, and looks instead at the classes of lawful and of outlawed events.

## B. Big and Little Miracles

It has often been suggested, not often by well-wishers, that I should distinguish big and little miracles thus: big miracles are other-worldly events that break many of the laws that actually obtain, whereas little miracles break only a few laws. I think this proposal is thoroughly misguided. It is a good thing that I never endorsed it, and a bad thing that I am sometimes said to have endorsed it.

Consider two cases. (1) By "laws" we might mean *fundamental* laws: those regularities that would come out as axioms in a system that was optimal among true systems in its combination of simplicity and strength. If the hopes of physics come true, there may be only a few of these fundamental laws altogether. Then *no* miracle violates many fundamental laws; *any* miracle violates the Grand Unified Field equation, the Schrödinger equation, or another one of the very few, very sweeping fundamental laws.

Or (2) by "laws" we might rather mean fundamental *or derived* laws: those regularities that would come out as axioms *or theorems* in an optimal system. Then any miracle violates infinitely many laws; and again it doesn't seem that big miracles violate more laws than little ones.

It's a blind alley to count the violated laws. What to do instead?

end p.55

Take the laws collectively; distinguish lawful events from unlawful ones. (For example, lawful pair-annihilations with radiation from unlawful quiet disappearings of single particles without a trace.) In whatever way events can be spread out or localized, unlawful events can be spread out or localized. In whatever way several events can be alike or varied, several unlawful events can be alike or varied. In whatever way we can distinguish one simple event from many simple events, or from one complex event consisting of many simple parts, we can in particular distinguish one simple unlawful event from many, or from one complex event consisting of many simple unlawful parts. A big miracle consists of many little miracles together, preferably not all alike. What makes the big miracle more of a miracle is not that it breaks more laws; but that it is divisible into many and varied parts, any one of which is on a par with the little miracle.

## C. Worlds to Which Convergence Is Easy

Begin with our base world $w_0$, the deterministic world something like our own. Proceed to $w_1$, the world which starts out just like $w_0$, diverges from it by a small miracle, and thereafter evolves in accordance with the laws of $w_0$. Now extrapolate the later part of $w_1$ backward in accordance with the laws of $w_0$ to obtain what I shall call a *Bennett world*.[3] This Bennett world is free of miracles, relative to $w_0$. That is, it conforms perfectly to the laws of $w_0$; and it seems safe to suppose that these are the laws of the Bennett world also. From a certain time onward, the Bennett world and world $w_1$ match perfectly, which is to say that $w_1$ converges to the Bennett world. Further, this convergence is accomplished by a small miracle: namely, the very same small miracle whereby $w_1$ diverges from $w_0$. For we had already settled that this small divergence miracle was the only violation by $w_1$ of the laws of $w_0$, and those are the same as the laws of the Bennett world. Thus the Bennett world is a world to which convergence is easy, since $w_1$ converges to it by only a small miracle.

What then becomes of my asymmetry of miracles? I said that

end p.56

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

"divergence from such a world as $w_0$ is easier than perfect convergence to it. Either takes a miracle . . . but convergence takes very much more of a miracle." To be sure, I said that it might be otherwise for a different sort of world. But the Bennett world seems to be a world of the same sort as $w_0$. After all, it has the very same laws.

No. Same laws are not enough. If there are *de facto* asymmetries of time, not written into the laws, they could be just what it takes to make the difference between a world to which the asymmetry of miracles applies and a world to which it does not; that is, between a world like $w_0$ (or ours) to which convergence is difficult and a Bennett world to which convergence is easy. Consider, for instance, Popper's asymmetry.[4] That is not a matter of law, so it could obtain in one and not the other of two worlds with exactly the same laws. Likewise in general for the asymmetry of overdetermination.

A Bennett world is deceptive. After the time of its convergence with $w_1$, it contains exactly the same apparent traces of its past that $w_1$ does; and the traces to be found in $w_1$ are such as to record a past exactly like that of the base world $w_0$. So the Bennett world is full of traces that seem to record a past like that of $w_0$. But the past of the Bennett world is not like the past of $w_0$: under the laws that are common to both worlds, the past of the Bennett world predetermines that Nixon presses the button, whereas the past of $w_0$ predetermines that he does not. Further, we cannot suppose that the two pasts are even close. As I noted in discussing world $w_2$, there is no reason to think that two lawful histories can, before diverging, remain very close throughout a long initial segment of time. To constrain a history to be lawful in its own right, and to constrain it also to stay very close to a given lawful history for a long time and then swerve off, is to impose two very strong constraints. There is not the slightest reason to think the two constraints are compatible.

To be sure, any complete cross section of the Bennett world, taken in full detail, is a truthful record of its past; because the Bennett world is lawful, and its laws are *ex hypothesi* deterministic (in both directions), and any complete cross section of such a world is lawfully sufficient for any other. But in a world like $w_0$, one that manifests the ordinary *de facto* asymmetries, we also have plenty of very *incomplete* cross sections that postdetermine incomplete cross sections at earlier times. It is these incomplete postdeterminants that are missing from

end p.57

the Bennett world. Not throughout its history; but the postdetermination across the time of convergence with $w_1$ is deficient.

Popper's pond is deceptive in just the same way. Ripples rise around the edge; they contract inward and get higher; when they reach the center a stone flies out of the water—and then the pond is perfectly calm. What has happened is the time-reversed mirror image of what ordinarily happens when a stone falls into a pond. It is no less lawful; the violated asymmetries are not a matter of law. There would be no feasible way to detect what had happened. For there would be no trace on the water of its previous agitation; and the rock would be dry, the air would bear no sound of a splash, the nearby light would bear no tell-tale image, . . . . In short, a perfect cover-up job—and without any miracle! But not in a world like $w_0$, and not in a world like ours. To be sure, if the laws are deterministic, the event is postdetermined by any complete cross section afterward. But we lack the usual abundance of lesser postdeterminants.

## D. The Indeterministic Case

I assumed determinism for the sake of the argument. I considered the deterministic case in order to oppose the view that the asymmetries under consideration arise out of one-way indeterministic branching.

That is not to say, of course, that I assume determinism *simpliciter*. I. do not. Accepted physics, after all, is not deterministic. It is hard to know what to make of the indeterminism in present-day quantum mechanics. *Pace* Einstein, indeterminism *per se* is credible enough. But the trouble is that the only indeterministic process in nature—reduction of the wave function, as opposed to Schrödinger evolution—is supposed to be special to the phenomenon of measurement.[5] And the

end p.58

idea that a unique microphysical process takes place when a person makes a measurement seems about as credible as the idea that a unique kind of vibration takes place when two people fall truly in love. Instrumentalist philosophy among physicists doesn't help matters, though perhaps the quantum theory of measurement is such a disaster that it *deserves* to be dismissed as a mere instrument. Which parts of present theory are fact, which fiction? What will remain when the dust settles?

I can only guess; my guess is not especially well informed; but for what it is worth, I guess as follows. The theoretical foundation of quantum mechanics is probably wrong to say that reduction is brought on when people measure. But the working quantum mechanics of

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

radioactive decay, coherent solids, chemical bonding, and the like can somehow stand on its own. It does not need this unfortunate anthropocentric foundation.[6] Then the laws of nature that govern our world really are indeterministic. Whatever we make of the reduction of the wave function supposedly brought on by measurement, at any rate there are chance processes involved in radioactive decay, in the making and breaking of chemical bonds, in ionization, in the radiation of light and heat, and so on. These processes are pervasive. So much so that not only is the world as a whole indeterministic, but also it can contain few if any deterministic enclaves.

If so, then what becomes of my asymmetries? In one way, the problem is easier. Divergence no longer requires a small miracle, not if there are abundant opportunities for divergence in the outcomes of chance processes. (If indeterministic processes were very scarce, miracles might still be required sometimes. But that is probably not the case for our world.) So in the indeterministic case it does not matter whether I am right to count small miracles as relatively cheap dissimilarities. Our divergences come cheaper still.

The thing to say about approximate convergence remains the same. Even if approximate convergence is cheap—and even if it is cheaper still when it can be had without even a little miracle—still we can say that it counts for little or nothing, so it is not so that if Nixon had

end p.59

pressed the button there would have been approximate convergence to our world, and no holocaust.

But what about perfect convergence? Here, indeterminism makes my problem harder. It is not to be said that the similarity achieved by perfect convergence counts for little or nothing. For it is just like the past similarity that has decisive weight in the deterministic case, tilting the balance in favor of last-minute divergence instead of difference throughout the past. I said that perfect convergence would take a big, widespread, varied miracle—a miraculously perfect cover-up job. But if chance processes are abundant, as I have guessed that they are, why couldn't they accomplish the cover-up? Why couldn't convergence happen without any miracles at all, simply by the right pattern of lawful outcomes of many different chance processes? Call such a pattern a *quasi-miracle*. It is extraordinarily improbable, no doubt, but it does not violate the laws of nature that prevail at our world.

What must be said, I think, is that a quasi-miracle to accomplish perfect convergence, though it is entirely lawful, nevertheless detracts from similarity in much the same way that a convergence miracle does. That seems plausible enough. (Though the test of the hypothesis is not in its offhand plausibility, but its success in yielding the right counter-factuals.) The quasi-miracle would be such a remarkable coincidence that it would be quite unlike the goings-on we take to be typical of our world. Like a big genuine miracle, it makes a tremendous difference from our world. Therefore it is not something that happens in the closest worlds to ours where Nixon presses the button. These worlds have no convergence miracles, and also no convergence quasi-miracles. So the case turns out as it should: the closest worlds where Nixon presses the button are worlds where a holocaust ensues.

My point is not that quasi-miracles detract from similarity because they are so very improbable. They are; but ever so many unremarkable things that actually happen, and ever so many other things that might happen under various counterfactual suppositions, are likewise very improbable. What makes a quasi-miracle is not improbability *per se*, but rather the remarkable way in which the chance outcomes seem to conspire to produce a pattern. If the monkey at the typewriter produces a 950-page dissertation on the varieties of anti-realism, that is at least somewhat quasi-miraculous; the chance keystrokes happen to simulate the traces that would have been left by quite a different process. If the monkey instead types 950 pages of jumbled letters, that is not at all quasi-miraculous. But, given suitable assumptions about what sort of chance device the monkey is, the one text is exactly as

end p.60

improbable as the other. (It is irrelevant to compare the probability that there will be *some* dissertation with the probability that there will be *some* jumble—the monkey does not just select one or the other kind of text, but also produces a particular text of the selected kind.) The pattern of systematic falsification of traces required for perfect convergence is quasi-miraculous in the same way.

(What if, contrary to what we believe, our own world is full of quasi-miracles? Then other-worldly quasi-miracles would not make other worlds dissimilar to ours. But if so, we would be very badly wrong about our world, so why should we not turn out to be wrong also about which counterfactuals it makes true? I say that the case needn't worry us. Let it fall where it may.)

In the deterministic case, the asymmetry of counterfactuals derives from an asymmetry of miracles: divergence takes less of a miracle than (perfect) convergence. Likewise in the indeterministic case we have an asymmetry of quasi-miracles. Convergence to an indeterministic world of the sort that ours might be takes a quasi-miracle; divergence from such a world does not. (I do not speak of small quasi-miracles; what corresponds to a small miracle in the deterministic case is a perfectly commonplace chance occurrence.) The asymmetry is made plausible by the same thought-experiment as before: think, in some detail and without neglecting imperceptible differences, of what would be needed for a perfect cover-up.

The trouble with using quasi-miracles as a weighty respect of dissimilarity is that it seems to prove too much, more than is true. For if

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

quasi-miracles make enough of a dissimilarity to outweigh perfect match throughout the future, and if I am right that counterfactuals work by similarity, then we can flatly say that if Nixon had pressed the button there would have been no quasi-miracle. But if chance processes are abundant, and would have been likewise abundant if Nixon had pressed the button, then it that case there would have been *some* chance of a quasi-miracle. To be sure, the probability would have been very low indeed. But it would not have been zero.

But if there would have been some minute probability of a quasi-miracle, does it not follow that there might have been one? And if there might have been one, then is it not false to say that there would not have been one? True, it would have been overwhelmingly probable that there not be one. But may we say flatly that this improbable thing would not have happened?

(Note that I am not talking about probabilities that certain counterfactuals are true. Rather, the consequents of the counterfactuals have to

end p.61

do with probabilities. In particular, they have to do with objective single-case chances, as of the time right after the hypothetical pressing, of the patterns of events that would comprise a suitable quasi-miracle.[7] )

Is there, perhaps, an exact balance? Suppose that perfect match throughout the future contributes to similarity exactly as much as the quasi-miracle needed to achieve that match detracts from similarity. Then worlds with a quasi-miraculous convergence have no net advantage, and no net disadvantage. Then they can be some, but not all, of the closest worlds where Nixon pressed the button. That seems to give the right counterfactuals: it is not so that if he had pressed the button then there *would* have been quasi-miraculous convergence, and such convergence would not have been at all probable; but it is so that if Nixon had pressed the button then there *might* have been quasi-miraculous convergence. So far, so good. But this solution (besides seeming artificial) fails to solve the whole problem. What about other quasi-miracles: patterns of outcomes of chance processes that are just as much remarkable coincidences, just as improbable, just as dissimilar from what typically goes on at our world—but do not yield convergence? On the balance hypothesis, these *non-convergence quasi-miracles* detract greatly from similarity and bring no compensating gain. So they, unlike the convergence quasi-miracles, are not to be found at any of the closest worlds where Nixon had pressed the button. And that seems wrong. It seems that we should say the same thing about *any* quasi-miracle, whether or not it yields convergence: if Nixon had pressed the button, it would have had some minute probability of happening, hence if so it might have happened, hence we should not say flatly that it would not have happened. So the hypothesis of exact balance does not save the day and I am still in trouble.

The line of retreat, of course, is asymmetry by fiat. Analysis 1, which drops the whole idea that counterfactuals work by similarity, is still available. It has no need of determinism. Or we could complicate the weighting of respects of similarity so that perfect match in the past weighs heavily but perfect match in the future counts for nothing. (More precisely: perfect match before and after the time relevant to the counterfactual supposition in question—as it might be, the time of Nixon's supposed pressing of the button.) Either way, we build an asymmetry between the directions of time into our very analysis: counterfactual, and hence on my view causal, dependence just *consists*

end p.62

in part of temporal order. I still say that won't do. It imposes *a priori* answers on questions that ought to be empirical. No; the asymmetry of counterfactual dependence should come from a symmetrical analysis and an asymmetrical world.

What is to be done? Our trouble was caused by an apparent logical connection between counterfactuals about what would happen, counterfactuals about what might happen, and counterfactuals about what the chances would be. One escape route is to reconsider that connection. Indeed, the connection seemed intuitively right, and I would be reluctant to challenge it just as a cure for my present trouble. But it needs challenging also for other reasons.

Recall the problem. By treating quasi-miracles as a weighty respect of dissimilarity, I make it turn out that there are no quasi-miracles of any kind, and hence there is no quasi-miraculous convergence, at any of the most similar worlds where Nixon pressed the button. That means that:

(1) If Nixon had pressed the button, there would not have been a quasi-miracle.

But quasi-miracles are just certain special patterns of outcomes of chance processes, and the chances would have been much the same if Nixon had pressed the button. That means that:

(2) If Nixon had pressed the button, there would have been some minute chance of a quasi-miracle.

We had better accept both (1) and (2). But they seem to conflict. Or do they? Considered by themselves, there is no very clear impression of conflict. Above, to create a semblance of conflict, I went in two steps, by way of:

(3) If Nixon had pressed the button, there might have been a quasi-miracle.

Whether or not (1) and (2) conflict, it certainly seems that (1) and (3) conflict, and it also seems that (2) implies (3). But perhaps we are being fooled by an ambiguity in (3).

I have hitherto advocated a “not-would-not” reading of “might” counterfactuals, on which (3) comes out as:

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

(3-nwn) It is not the case that: if Nixon had pressed the button, there would not have been a quasi-miracle.

end p.63

But perhaps there is also a "would-be-possible" reading, on which (3) comes out as:

(3-wbp) If Nixon had pressed the button, it would be that: a quasi-miracle is possible.

The readings differ as follows. (3-nwn) means that some of the most similar worlds where Nixon pressed the button are worlds where a quasi-miracle happens; whereas (3-wbp) means that all of them are worlds where it is possible for a quasi-miracle to happen. If all of them are worlds where there is an unfulfilled possibility of a quasi-miracle, that makes (3-nwn) false and (3-wbp) true. And if we take possibility to mean non-zero chance (as of the time of the pressing), then that is exactly the situation that makes (1) and (2) both true together. Indeed, (1) conflicts with (3-nwn); indeed, (2) implies (3-wbp); but (1) and (2) are compatible.[8]

I note that the same problem arises in consequence of my treatment of counterfactuals with true antecedents. Suppose that our world is an *A*-world with an unfulfilled non-zero chance of *B*. Then, since a counterfactual with a true antecedent is true iff its consequent is,[9] we have a pair of true counterfactuals that parallel (1) and (2).

(4) If it were that *A*, then it would be that not *B*.

(5) If it were that *A*, then there would be some chance that *B*.

Thus (4) and (5), on my account, are compatible. Yet they may appear to conflict if we consider:

(6) If it were that *A*, then it might be that *B*.

This counterfactual seems to conflict with (4) and to be implied by (5). I reply that on the "not-would-not" reading (6) conflicts with (4) and is false, whereas on the "would-be-possible" reading it is implied by (5) and is true.

end p.64

In fact, our problem is more far-reaching still. If we want any kind of similarity theory of counterfactuals, we dare not treat "there would be some chance of it" and "it would not happen" in general as incompatible. Suppose for *reductio* that counterfactuals of these two kinds are in general incompatible. Let *C* be any proposition that might obtain or not as a matter of chance; let *u* and *v* be a *C*-world and a not-*C*-world, respectively, but let them both be worlds that have a chance of going either way; let *A* be the proposition that holds at these two worlds, and no others; and let *w* be any third world. It is true at *w* that if *A*, there would be some chance that *C*; so by the supposed incompatibility, it is false at *w* that if *A*, it would be that not-*C*; so *u* must be at least as close to *w* as *v* is. Likewise, putting not-*C* in place of *C, v* must be at least as close to *w* as *u* is. That is, worlds *u* and *v* are tied in closeness to any third world. But *u* and *v* were *any* two worlds that differ in respect of the outcome of a matter of chance—no matter how much they may differ in other ways as well! This completes the *reduction*.[10]

We can have a simpler *reductio* if we suppose that it is legitimate to mention chances in the antecedent of a counterfactual—and how can that fail to be legitimate, if chances are indeed an objective feature of the world? What would be the case if there were an unfulfilled chance of *C*? If so, then there would be some chance that *C*. But if so, then also it would not be that *C*. So here we have a counterexample to the supposed incompatibility, just on the principle that a counterfactual holds when the antecedent implies the consequent.

So the supposed incompatibility had better be rejected. The reconciliation of (1) with (2), (4) with (5), and the like is by no means just a dodge to defend my controversial views about time's arrow and about counterfactuals with true antecedents. But it does serve, *inter alia*, to defend them. We can count quasi-miracles as weighty dissimilarities from actuality; we can persuade ourselves by examples that perfect reconvergence to a world like ours would require, if not a big miracle, at least a quasi-miracle; we can conclude that if Nixon had pressed the button, there would have been no perfect convergence; and still we can say, as we should, that there would have been some minute chance of perfect convergence.

end p.65

## E. Ubiquitous Traces and Common Knowledge

My argument for an asymmetry of miracles (or of quasi-miracles) relied on an empirical premise: at a world like ours, everything that happens leaves many and varied traces, so that it would take a big miracle—equivalently, many and varied small miracles working together—to eradicate those traces and achieve reconvergence. But I need more than merely the truth of that premise. I need common knowledge of it. For if the premise were true but generally disbelieved, and if our counterfactuals work as I say they do, then we ought to find people often accepting the counterfactuals that would be true on my account if that premise were false. We ought to find them saying that if Nixon had pressed the button, the future would have been no different, there would have been convergence and no holocaust. In illustrating the multitude of traces that the pressing would have left, and the difficulty of a perfect cover-up, I relied on a certain amount of

**Philosophical Papers Volume II**
**Lewis, David,** Professor of Philosophy, Princeton University
Print publication date: 1987, Published to Oxford Scholarship Online: November 2003
Print ISBN-13: 978-0-19-503646-6, doi:10.1093/0195036468.001.0001

scientific knowledge that many people do not share. I may have explained why the right counterfactuals come out true according to my beliefs. But I have done nothing to explain why ignorant folk accept those same counterfactuals.

I reply that everyone believes in ubiquity of traces. Maybe not everyone can illustrate the point in the way I did (though I must say that I did not use anything very esoteric) but they can still think that *somehow* everything leaves many and varied traces.

Consider detective stories. Seldom are they written by, or for, expert scientists. The background against which they are to be read is common knowledge, not expert knowledge. And part of that background is the assumption that events leave many and varied traces. Else the plots would not make sense. We are supposed to marvel at the skill of the detective in spotting and reading the traces. We are not supposed to marvel that the traces are there at all. Ignorant or expert, anyone knows better than to read the tale as a hard-luck story: how the criminal was caught because he was especially unfortunate in leaving traces. And anyone knows better than to read the tale as science fiction: how things would be in a bizarre world where things leave far more traces than they do in ours. No; it is supposed to be a tale of a world like ours, and the ubiquity of traces is part of the likeness.

end p.66