Counterfactuals and Explanation

BORIS KMENT

On the received view, counterfactuals are analyzed using the concept of closeness between possible worlds: The counterfactual 'If it had been the case that p, then it would have been the case that q' is true at a world w just in case q is true at all the possible p-worlds closest to w. The degree of closeness between two worlds is usually thought to be determined by weighting different respects of similarity between them. The question I consider in the paper is which weights attach to different respects of similarity. I start by considering Lewis's answer to the question and argue against it by presenting several counterexamples. I use the same examples to motivate a general principle about closeness: If a fact obtains in both of two worlds, then this similarity is relevant to the closeness between them if and only if the fact has the same explanation in the two worlds. I use this principle and some ideas of Lewis's to formulate a general account of counterfactuals, and I argue that this account can explain the asymmetry of counterfactual dependence. The paper concludes with a discussion of some examples that cannot be accommodated by the present version of the account and therefore necessitate further work on the details.

On the received view, counterfactuals are analyzed in terms of a relation of similarity (or 'closeness') between possible situations. The idea is neatly expressed in the opening sentence of David Lewis's book on the topic:

> "*If kangaroos had no tails, they would topple over*" seems to me to mean something like this: in any possible state of affairs in which kangaroos have no tails, and which resembles our actual state of affairs as much as kangaroos having no tails permits it to, the kangaroos topple over. (Lewis 1973, p. 1)

When developed formally, the account is formulated in terms of maximal possible states of affairs, or possible worlds. Certain technical niceties aside, the theory assigns the following truth-conditions to counterfactuals:

(0) 'If it had been the case that *A*, then it *would* have been the case that *C*' is true just in case *C* is true in *all* the possible *A*-worlds closest to the actual world.

'If it had been the case that *A*, then it *might* have been the case that *C*' is true just in case *C* is true in *some* of the possible *A*-worlds closest to the actual world.[1]

What I have described so far is merely a framework for an account of counterfactual conditionals. More needs to be said about the relation of closeness or similarity that enters into this account. It has frequently been noted that this concept of similarity cannot be the one we use in our offhand judgments about the overall similarity between worlds.[2] In order to see this, consider the counterfactual

(1) If Nixon had pressed the button, there would have been a nuclear catastrophe.

We may believe that this sentence is true. But offhand it might seem that a world in which a nuclear catastrophe ensues after the button pressing is very unlike our world. In that world, the earth is devastated, in ours it is not. The nuclear-disaster world might seem much less similar than another world in which Nixon presses the button but the signal dies in the wire that leads from the button to the launch pad, so that the rest of history is very similar to that of our world. This seems to show that, if we used our offhand judgments about similarity to give an account of the truth-conditions of counterfactuals, we would get the wrong results. We would have to say that, if Nixon had pressed the button, everything would have been fine.

Given that the standards of similarity that are relevant to the truth-conditions of counterfactuals cannot be those that govern the familiar notion of offhand similarity,

[1] The theoretical framework described is due to Stalnaker (1968) and Lewis (1973a). Other significant work done in that framework includes Jackson (1977), Bennett (1984), and Lewis (1979), (1986a), among others.

[2] See, for example, Bennett (1974), Fine (1975), and Lewis (1979), pp. 41-48. Also cp. Lewis (1973), p. 76. The example is a variant of Fine's.

it is urgently necessary to give some account of them. It needs to be specified in detail which respects of similarity between two worlds are relevant to the degree of closeness between them, and how weightily each of them contributes to determining the degree of closeness.

We should not assume that there is *one* relation of similarity that is relevant to all possible uses of counterfactuals. In fact, it is commonly believed that the standards of similarity vary across different contexts of use. In order to see what motivates this assumption, consider an example due to Jackson (1977, p. 9): Frank is in a room on the tenth floor of a building. There are no nets or other contraptions to break the fall of someone jumping out of the window onto the street below. It seems that we can safely say that Frank would get badly hurt if he were to jump out of the window. But Frank, on hearing this conclusion, might reply: 'I'm a sensible fellow. I would never jump out of a tenth-floor window, unless I had made sure that there was a safety net. So, if I were to jump, a net would be in place, and I would be fine.' Frank's reasoning might convince us of the truth of his counterfactual. And yet his conditional seems to be incompatible with the one we endorsed before. The most obvious diagnosis is that the truth-conditions of counterfactuals are context-dependent. In some contexts, worlds in which Frank jumps despite the absence of a net count as closer than worlds in which he first places a net below the window and then jumps. In other contexts, it is the other way around. In contexts of the first sort, we can say that Frank would get injured if he were to jump. In those of the second type, we ought to say that he would be fine.

Examples of this kind convinced Lewis of the context-dependence of counterfactuals. But Lewis also believed that there is a default assignment of truth-conditions to counterfactuals, an assignment we choose when interpreting the utterance of a counterfactual unless our presumption in favour of it is cancelled by distinctive features of the context (1979, p. 34–5.). This seems plausible enough in the example of the last paragraph: If presented with the case out of the blue and asked for a judgment, we would say that Frank would get badly hurt if he were to jump. It requires some stage-setting (like that provided by Frank's utterance) to create a context in which we are willing to agree that he would have been fine.

For reason on which I cannot expand here, I think that a similarity account like (0) is a little bit too simple, but I think that it is a good approximation to the truth, and it can serve as a perfectly good working account for my purposes. I also believe that the truth-conditions of counterfactuals are context-dependent, and I accept Lewis's view that there is a default assignment of truth-conditions. (These will be called the 'standard truth-conditions'.) These standard truth-conditions might be vague, and it could be that this vagueness is resolved, if at all, only by further aspects of the context. (That a context calls for the standard assignment of truth-conditions to a certain counterfactual might not be the only feature of that context that contributes to determining the truth-conditions of the counterfactual.[3])

I will propound a certain general principle about the closeness relation that enters into the standard truth-conditions. I will try to make it plausible that this principle can explain a variety of data, including the temporal asymmetry of counterfactual dependence. I believe that the principle should be at the centre of any account of the truth-conditions of counterfactuals. (As I note in section 9, I think that the explanatory power of the principle extends further than I will be able to describe in this paper, and I expand on it in my (forthcoming).) The formulation of the principle I will offer is merely intended as an approximation. As I indicate in section 8, I think that more work on the details is needed. But this is a task for another occasion. In this paper my concern will be with the larger picture.

1. The intuitive data

Attempts to specify the closeness relation that matters to the truth-conditions of standard counterfactuals typically start by considering a special case: Counterfactuals whose antecedents are nomically possible and deal with matters of particular local fact. Intuition seems to furnish two important data about counterfactuals of this kind, stated in (2) and (3) below.

[3] Consider a standard example, the pair of conditionals 'If Caesar had been in command in Korea, he would have used the A-bomb', and 'If Caesar had been in command in Korea, he would have used catapults'. It is far from obvious that the default interpretation of these counterfactuals tells us for each of them whether it is true or false. It is perhaps more plausible to think that, if there is a default interpretation of counterfactuals, the truth-conditions it assigns must be vague enough to leave open the truth-values of the two conditionals at issue.

(2) *Counterfactual dependence is temporally asymmetrical*. If matters of particular local fact at one time had been different, then things later on would have been different as well; but earlier matters would have been pretty much the way they actually were. If Nixon had pressed the button, then a day later the world would have been radically different from what it was actually like. But until shortly before the button-pressing, matters would have been pretty much the way they actually were.

(3) *Conformity to the laws of the actual world contributes to the closeness of an antecedent-world.* If Nixon had pressed the button, then events would still have tended to conform to the actual laws of nature. In particular, if the missile system is set up in such a way that the only lawful course of events that can follow the pressing of the button leads to a nuclear explosion, then there is a nuclear catastrophe in all the closest antecedent-worlds.

The special role of the laws in the truth-conditions of counterfactuals was already emphasized by Goodman in his classic paper on counterfactuals (1947), and has frequently been highlighted in subsequent theories of counterfactuals. In fact, the ability to support counterfactuals has been regarded as one of the most important features that distinguish laws from accidental universal generalizations.

According to (2), the closest antecedent-worlds should, all other things being equal, be like our world until (shortly before) the antecedent-time. According to (3) they should, *ceteris paribus*, conform to the actual laws. Now suppose that determinism is true, in the sense that any two possible worlds that perfectly conform to the actual laws of nature and which are perfectly alike throughout some extended initial segment of their histories are perfectly alike throughout their histories. If the

antecedent is false, then under determinism the degree to which a possible antecedent-world conforms to one of the two *desiderata* limits the degree to which it can conform to the other. For if determinism is true, then every initial segment of the history of the actual world, together with the laws, determines that the antecedent is false (in the sense that the antecedent is false in any possible world that is like our world throughout this initial segment and conforms perfectly to the actual laws thereafter).[4] This implies that no possible antecedent-world can perfectly conform to the actual laws *and* be like our world until shortly before the antecedent-time.

Philosophers differ in the way they prefer to respond to this finding. Some theories, like that which Jonathan Bennett propounded in his (1984) (and which Bennett himself criticized later on (2001, 2003, section 80)), stress the importance of conformity to the actual laws, at the expense of similarity in pre-antecedent matters. On Bennett's 1984 account, if the antecedent is consistent with all the laws, then the closest antecedent-worlds conform perfectly to the actual laws. Under determinism, this means that (if the antecedent is false, then) the closest antecedent-worlds differ from our world throughout the pre-antecedent time. Lewis places greater emphasis on pre-antecedent match, at the price of small violations of the actual laws: Consider a counterfactual whose antecedent is false and deals with matters of particular local fact, such as (1). Lewis suggests that the closest antecedent-worlds are exactly like the actual world until shortly before the button-pressing. After that they diverge from ours just enough to allow Nixon to press the button. The transition from the actual past to a course of events that makes the antecedent true occurs in some smooth way, without abrupt discontinuities. Suppose that in our world, Nixon was on the second floor at *t* and that the button is on the first floor. In that case, if Nixon had pressed the

[4] In Sct. 5, I will suggest that laws can have exceptions, in the sense that a universal generalization L can be a law in a possible world *w* even though there are exceptions to L in *w*. If we accept that laws can have exceptions in this sense, then we need to qualify the claim that under determinism every initial segment of the history of the world, together with the laws, determines that the antecedent is false. It could be that some initial segment of the history of the world, together with the laws, determines that the antecedent is *true* (in the sense that the antecedent is true in every possible world that is just like our world throughout this initial segment and which perfectly conforms to the actual laws thereafter), but that after the end of this initial segment some law is violated in our world, so that the antecedent turns out false anyway. However, even if laws can have exceptions, it *may* still hold for many (or even all) false antecedents that every initial segment of the history of the universe, together with the laws, determines that the antecedent is false, so that the tension between the two *desiderata* stated in (2) and (3) still arises in many cases.

button at *t*, he would first have descended to the first floor, by taking the stairs or the elevator, in order to get the button within the reach of his fingertips. So, the closest antecedent-worlds must be ones that diverge from ours shortly before *t* so as to allow Nixon to make his way to the button. If our world is deterministic, then the divergence of the antecedent-world from our world requires some infringement of the laws of our world, a small 'miracle,' as Lewis aptly calls it. Some very small and inconspicuous such violation will be enough to bring about the needed divergence: Perhaps some extra neurons fire in Nixon's brain. After the button-pressing, the closest antecedent-worlds evolve in accordance with the laws of our world. If the missile system is absolutely reliable, the nuclear catastrophe ensues, and the counterfactual comes out true.

I will call antecedent-worlds of the kind described—those that are exactly like our world until shortly before the antecedent-time, then smoothly diverge just enough to make the antecedent true, and afterwards evolve in accordance with the actual laws—the 'well-behaved antecedent-worlds.' Lewis's idea, then, is that something like the following is more or less true of counterfactuals whose antecedents are falsehoods about matters of particular local fact and are consistent with all the actual laws:

(4) The well-behaved antecedent-worlds are closer than any other antecedent-worlds.

As David Lewis notes (1979, pp. 38–41), (4) cannot serve as an account of closeness as it stands, for at least the following two reasons: *First*, (4) is insufficiently general. It applies only to counterfactuals whose antecedents deal with matters of particular local fact and are consistent with the laws. It leaves conditionals like the following unaccounted for:

If there were cricket players moving faster than light, then … .
If there were no forces of gravitation, then … .

Secondly, (4) is insufficiently flexible. It lays down that, for *every* counterfactual whose antecedent is about matters of particular local fact (and is consistent with the laws), the closest antecedent-worlds are just like our world until shortly before the antecedent-time. It thus determines that, for any actual localized matter of particular fact, if that matter of fact had not obtained, all matters until shortly before it would have been just the same (though it leaves open the possibility that matters later on would have been quite different). Now, it seems admittedly right in most ordinary-life cases that the past is more or less counterfactually independent of the present. However, if we made (4) part of our account of the truth-conditions of counterfactuals, we would have to regard the past's counterfactual independence of the present as a necessary truth, and could not leave room for possible exceptions to it, even in the most outlandish circumstances. But, Lewis thinks, and I agree, that it would be rash to think that backward counterfactual dependence is metaphysically impossible. Cases of backward causation, as in precognition and time travel, may be metaphysically possible, and in such cases even the distant past might counterfactually depend on the present. (I have built a time machine, which has a dial that I can set to the time I want to travel to. I set it to 1600, get into the machine and travel to the year 1600. It seems right to say that I would instead have arrived in 1500 if I had set the dial accordingly.) It therefore seems that, if the asymmetry of counterfactual dependence is so common in everyday life, then this might be due, not merely to the truth-conditions of counterfactuals, but also, in part, to the scarcity of time travel and kindred phenomena. A good account of counterfactuals should not write it into their truth-conditions that the past is counterfactually independent of the present. Instead, it should allow us to *explain* this fact by appealing both to certain aspects of the truth-conditions of counterfactuals and to certain features of the world.

Despite the shortcomings of (4) as a general account of counterfactuals, Lewis thinks that it assigns the right truth-values to most ordinary-life counterfactuals whose antecedents deal with matters of particular local fact. He therefore assumes that a good general theory of counterfactuals should agree with (4) throughout a considerable range of cases.

Lewis's goal, then, is to find a theory of the standard truth-conditions of counterfactuals that satisfies the following conditions: It applies to all counterfactuals, no matter what their antecedents are about. It does not write the asymmetry of counterfactual dependence into the account of the closeness relation, but allows us to explain it as being due, in part, to certain features of the world. And it agrees with (4) where (4) gets it right.

I will start my discussion by quickly considering the way Lewis tries to achieve this goal. I will then explain why I find his theory wanting. This discussion will set the stage for my own positive proposal.

2. Lewis's view

In presenting his theory, Lewis starts by formulating an account that is intended to assign the right truth-values to counterfactuals under determinism. He then modifies his theory to cover the indeterministic case as well. I, too, will follow this order of exposition. Consider example (1) and suppose that determinism is true and that the missile system is absolutely reliable.

As we have seen above, Lewis believes that an antecedent-world that diverges from our world shortly before the antecedent-time at the cost of a small miracle is closer than an antecedent-world that conforms perfectly to the actual laws but differs from our world throughout the pre-antecedent time. We can conclude that match in matters of particular fact throughout a massive region of space-time must contribute more to closeness than the avoidance of a small and inconspicuous miracle.

If this is so, then why do we not count as the closest antecedent-worlds those in which a small miracle *after* the antecedent-time prevents the nuclear disaster and thus ensures that the post-antecedent time is vastly more like the way it is in the actual world? Why not, for example, choose a world w_1 in which the electrical signal miraculously dies in the wire and no nuclear catastrophe ensues? Lewis notes that a small miracle of this kind would not lead to perfect reconvergence between w_1 and our world. In w_1, ever so many traces of Nixon's deed spread out through a vast portion of post-antecedent space-time: Nixon's finger leaves traces on the button, light waves travelling from Nixon's room into outer space bear images of his action,

and so forth. w_1 might be approximately like our world during the post-antecedent time, but not perfectly like it. And this, Lewis suggests, is why the post-antecedent similarities of w_1 do not counterbalance the small miracle they require: While a big space-time region of *perfect* match counts for more than the avoidance of a small miracle, a large spatio-temporal region of merely *approximate* match does not.

If extensive spatiotemporal regions of perfect match count for so much, then why do we not regard as the closest antecedent-worlds those in which *all* the traces of the button-pressing disappear, with the result of *perfect* reconvergence to the actual world? Under determinism, this requires a miracle. Now, as Lewis notes, no small and localized miracle can rid us of all the multifarious traces of Nixon's deed. The electrical signal must die in the wire, the wire has to cool down without heating up the insulation material around it, the images carried by the light waves need to vanish, and so on, and on, and on. The miracle needed to accomplish all that would be quite unlike the one required for the divergence from our world. It would be spread out through space and time and would involve many miraculous events of various different kinds. Big and widespread miracles of this sort detract too heavily from closeness to be counterbalanced even by massive gains in the size of the spatiotemporal region of perfect match.

These considerations suggest the following rules for weighting similarities:

> (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 spatiotemporal 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. (Lewis 1979, pp. 47–8)

So far we have centred on the deterministic case. But what if the actual world is indeterministic? Under indeterminism, perfect match until shortly before the antecedent-time still requires at most a small miracle,[5] and may not even require that (since the actual pre-antecedent history and the laws might leave open the possibility that Nixon presses the button). We can therefore explain in the same way as above why the closest antecedent-worlds are like our world until shortly before the antecedent-time. The treatment of approximate reconvergence can also remain the same: The imperfect post-antecedent match in an approximate-reconvergence antecedent-world contributes nothing to closeness. Hence, even if approximate reconvergence can be had without miracle (as may be the case under indeterminism), worlds with approximate reconvergence are no closer than those without.

The only issue that requires renewed attention is that of perfect reconvergence. Even under indeterminism antecedent-worlds that perfectly reconverge to our world are no closer than antecedent-worlds with no such reconvergence. (For we surely do not want to say that, if Nixon had pressed the button, then slightly later everything would have been just the way it actually was.) This cannot be explained by appealing to any big miracles needed for the reconvergence, since under indeterminism no miracle at all might be required. It may be that all that is necessary is that countless different chance processes come out just right to produce a pattern of particular facts just like the one we find in our world: By chance the images of the button-pressing disappear from the light rays travelling out of Nixon's room, the signal in the wire vanishes, as do Nixon's fingerprints, and so forth. Shortly after the button-pressing things look just the way they do in our world, and thus just as we would expect on the assumption that Nixon kept his fingers off the button.

Lewis uses the term 'quasi-miracle' for a combination of outcomes of random processes that produces a remarkable pattern that we would ordinarily take to be the outcome of a process of a quite different kind. (Another example of Lewis's involves

[5] Even under indeterminism a small miracle *might* be required for perfect match until shortly before the antecedent-time. If the world is indeterministic, then there might be forks, that is, cases in which the outcome of an indeterministic chance process determines which of several futures will be realized. But the thesis of indeterminism entails nothing about the frequency of forks, and it leaves open the possibility that they are extraordinarily rare. It might be that the latest fork before the antecedent-time is located a long time before the antecedent-time, so that any antecedent-world that is perfectly like our world until shortly before the antecedent-time contains a violation of law.

a monkey that produces a 950-page dissertation on anti-realism on a typewriter. The product of the process looks very much like the sort of thing usually produced by the ruminations of a graduate student.) Even under indeterminism, perfect reconvergence requires a quasi-miracle. And Lewis suggests that the occurrence of a quasi-miracle in an antecedent-world detracts from that world's closeness to ours as much as a big and widespread miracle does, so that a quasi-miracle cannot be counterbalanced even by perfect match throughout a massive chunk of space-time. This is why even under indeterminism the perfect-reconvergence world is less close to our world than an antecedent-world with nuclear catastrophe.[6]

In accordance with his goal, Lewis's account does not build the asymmetry of counterfactual dependence into the account of the standard truth-conditions of counterfactuals by fiat. Rather, it *explains* the asymmetry. The explanation appeals both to a certain feature of the world and to specific aspects of the standard truth-conditions of counterfactuals. The relevant feature of the world is the temporal asymmetry of miracles and quasi-miracles: A world's divergence from our world requires at most a small and inconspicuous miracle, while perfect convergence requires a big and complicated miracle or at least a quasi-miracle.[7] The relevant aspect of the standard truth-conditions of counterfactuals relates to the degrees to which quasi-miracles and different kinds of miracle detract from the closeness between worlds, and to the degrees to which different kinds of similarity in matters of particular fact contribute to closeness: Small miracles detract less from closeness than big miracles or quasi-miracles; a massive region of perfect match counts for enough to outweigh a small violation of law, though not a big and widespread violation or a

[6] Note that this result is stronger than Lewis needs. What Lewis set out to explain is why antecedent-worlds with perfect reconvergence are *no closer* than those without. His final result is that they are *less* close. In other words, Lewis commits himself to saying that, if Nixon had pressed the button, some things later on (say, a day later) would have been different from what they actually were. It is not intuitively obvious to me that this is true under indeterminism. In fact, I find it more plausible to say that, if our laws leave some chance that *all* traces of Nixon's deed disappear, then it is *not* true that there would have been no perfect reconvergence after the button-pressing. (Nor, of course, is it true that there *would* have been such a reconvergence.) There *might* have been, though it would have been extremely unlikely. Lewis discusses this intuition, and tries to explain it away ((1986a), pp. 61-5). I cannot embark on a discussion of his argument. But let me point out that the view I will propound can accommodate the intuition under consideration.

[7] This claim of Lewis's is controversial. See, e.g., Elga (2000).

quasi-miracle; approximate match counts for little or nothing and therefore cannot compensate even for a small miracle.

I take Lewis's discussion of the standard closeness relation to make three main contributions: *Firstly*, he proposes a response to the tension between two strictures on the range of closest antecedent-worlds, viz. the conformity to the actual laws and match in pre-antecedent matters: We ensure perfect pre-antecedent match until shortly before the antecedent-time by allowing for a small miracle. *Secondly*, he argues that the asymmetry of counterfactual dependence is not to be written into the truth-conditions of counterfactuals, but must be explained, in part, by appeal to features of the world. *Thirdly*, he sets out to formulate an account of counterfactuals that will allow him to give a suitable explanation of the asymmetry.

I have already indicated my appreciation of Lewis's second contribution. In section 5, I will discuss the first and present my reasons for agreeing with Lewis (I will describe a problem and offer a solution that requires me to agree with Lewis). But I will begin by considering Lewis's third contribution. I will argue that his explanation of the asymmetry of counterfactual dependence is mistaken. The findings that count against it will form the starting point of my discussion in the rest of the paper, and will suggest an alternative way of explaining the asymmetry of counterfactual dependence.

3. Closeness and causation

As we have seen, on Lewis's account, it 'is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.' And he adds in parentheses: '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' (Lewis 1979, p. 48). What Lewis says suggests that we might need to draw a distinction among the matters of particular fact in regions of imperfect match between those that matter to closeness and those that do not.

'Different cases come out differently,' Lewis says, and he has in mind two different kinds of example to be found in the literature.[8] Let us consider one case of each sort: In a certain world *w*, Bugsy has two indeterministic and fair coin-tossing devices, *A* and *B*. Each device, once activated, automatically tosses a coin after five minutes. Immediately before each toss, a random process is initiated inside the device and the outcome of this random process determines the outcome of the coin toss. Bugsy activates device *A*. Five minutes later, the coin is tossed and lands heads. Consider,

(5) If Bugsy had used device *B*, the coin would still have landed heads.

I think, as do most people I have asked, that this counterfactual is not true. If device *B* had been used, the coin might have landed heads, or it might have landed tails.

A little later in the history of the same world *w*, Bugsy again activates one of the coin-tossing devices and then offers you a bet on heads on the toss, but you decline it. Five minutes later, the coin is tossed and lands heads. Assume that your decision whether to accept the bet causally affects some of the goings-on inside the coin-tossing device: Your decision causes a certain utterance of yours, and the sound waves of this utterance penetrate the walls of the device and slightly change the distribution and motion of the air molecules inside it. However, the processes inside the device that are influenced by your decision do not in turn causally affect the outcome of the random process in any way. (This might not be compatible with the laws of the actual world, but it does not seem to be metaphysically impossible. I stipulate that it is compatible with the laws of the world *w* in which the coin toss takes place.) There is thus no causal connection whatsoever between your decision whether or not to accept the bet and the outcome of the coin toss. Now suppose that Bugsy says:

(6) If you had accepted the bet, you would have won.

[8] See, for example, Tichý (1976), Slote (1978), and Bennett (2003), Ch.15.

Most people I asked believe that Bugsy is right. I, too, feel inclined to agree with him. But if Bugsy is right, then it must be true that the coin would still have landed heads if you had accepted the bet.

In the closest worlds in which you accept the bet rather than to decline it, the traces that your decision leaves inside the box are different from what they are at *w*. Hence, in example (6) there is no perfect match between an antecedent-world and *w* with respect to the space-time region in which the random process and the coin toss take place. Similarly in example (5): In *w*, Bugsy uses coin-tossing device *A*, whereas in an antecedent-world he uses *B*. Hence, (assuming that there are differences between the two devices at least at the atomic level and between the distributions of air molecules inside them) there is no perfect match between *w* and the closest antecedent-worlds with respect to the space-time region of the random process and the coin toss.

As our intuitive judgments about the counterfactuals show, in example (6) well-behaved antecedent-worlds in which the coin toss has the same outcome as in *w* are closer than well-behaved antecedent-worlds in which it has a different outcome, despite the fact that the sameness of outcome does not increase the area of perfect match. By contrast, in example (5), some antecedent-worlds with a *different* outcome are among the closest antecedent-worlds. In other words, in the one case the similarity in outcome contributes to the closeness of an antecedent-world, while in the other case it does not.

It is not difficult to come up with an intuitively plausible explanation for this difference. The most natural diagnosis proceeds roughly along the following lines: Your decision whether or not to accept the bet does not make a difference to the outcome; that is: it does not *causally affect* the outcome. This is why we think that the outcome would have been just the same if you had made a different decision. Example (5) is different. If a different coin-tossing machine is used, then the causal history of the outcome of the coin toss is different (in the actual world certain processes involving the parts of machine *A* figure in the causal history of the outcome, whereas in a world in which machine *B* is used, its causal history instead

features certain processes involving the parts of *B*). Several authors who discuss pairs of examples of this kind provide diagnoses that are at least roughly along these lines.[9]

The above examples suggest, then, that similarities between two worlds *w* and *w** with respect to matters of particular fact concerning regions of approximate match contribute to the closeness between the worlds only if these matters of fact have the same causal history in the two worlds. How should we react to this discovery? Lewis's parenthetical remark quoted above already suggests a reaction: Refine the clause of his theory that relates to similarities in regions of approximate match by distinguishing between similarities concerning matters with the same causal histories and those concerning matters with different causal histories, and state explicitly that only the former contribute to closeness. This refinement of Lewis's account brings causal notions into the account of the standard closeness relation. Lewis, perhaps, would not have liked this, since he wanted to give an account of causation in terms of counterfactuals (1973b).[10] But for someone with no prior commitment to the counterfactual analysis of causation, the idea might seem attractive.

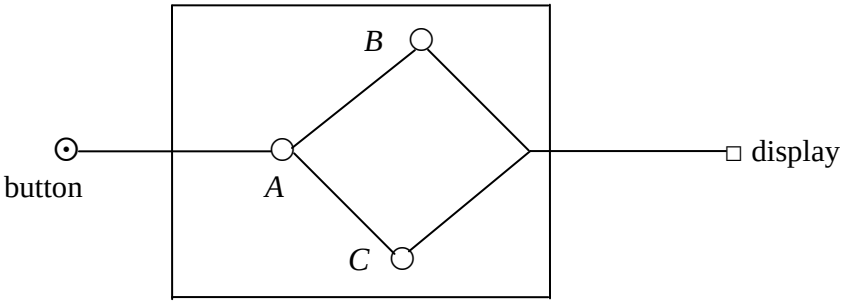
I do not think that this manoeuvre would solve the problem, however. For I think that the very phenomenon described above also arises for similarities in regions of *perfect* match: Such similarities are relevant to closeness only if they concern matters of particular fact with the same causal histories. Let me try to present a pair of examples that illustrates this. This pair of examples actually constitutes a counterexample to Lewis's theory. (The pair considered above does not. The above pair merely shows that some similarities concerning regions of imperfect match contribute to closeness while others do not. This is perfectly compatible with Lewis's account of counterfactuals.[11] For in his presentation of his account, Lewis explicitly

[9] Such a causal diagnosis of our intuitions about relevant examples was already given in Adams (1975), Ch. IV, Sct. 8 (in particular pp. 132f.), though it was not formulated in the closeness framework. Causal diagnoses formulated on the basis of the closeness account can be found, for example, in Mårtensson (1999); Edgington (2003); Bennett (2003), Ch.15; and Schaffer (2004). Also cp. Johnson (1991). The different causal diagnoses differ in matters of detail.

[10] See Edgington (2003) for a discussion of the incompatibility between the counterfactual account of causation and the causal diagnosis of our intuitive judgments about examples like the ones discussed above.

[11] There are other versions of example (6) in which the region of the coin toss in the closest antecedent-worlds is a region of *perfect* match (e.g. versions in which you are thousands of miles away from the room in which the coin is tossed and are watching the coin toss on television at the time the bet is

leaves open, and in fact suggests, that some similarities regarding regions of approximate match contribute to closeness while others do not.)



Consider a world *w* in which an indeterministic lottery draw takes place inside a box that contains the three random devices *A*, *B* and *C*. *A* is connected to a button outside the box. *A* is also linked by a wire to *B* and by another to *C*. *B* and *C*, in turn, are connected to a single other wire that leads to a display outside the box. When the button is pressed, an electrical signal travels into the box to *A*. When it reaches *A*, some random process inside *A* determines whether the signal travels on to *B* or to *C*. Once it arrives at *B* or *C*, it initiates another random process there that determines which ticket will win. The information about the outcome of the draw is then transmitted to the display outside the box. The random devices *B* and *C* give exactly the same chance to every possible outcome of the lottery. The interior of the box is causally isolated from its surroundings except for signals travelling into it from the button and signals that travel from it to the display. This might be incompatible with the laws of our world, but I stipulate that the laws of the world *w* (in which the draw takes place) allow it.

Suppose that the button is pressed. The signal travels to *A*, where it is determined that it will travel on to *B*. The random process inside *B* determines the outcome of the draw, and the result is transmitted to the display: Ticket number 17 has won. Assume further that during this entire period and throughout the rest of history, no causal signal passes into the box apart from the one coming from the button, and no causal

---

offered to you). These are still not counterexamples to Lewis' view, however. In the closest antecedent-worlds, in which the coin toss has the same result as in *w*, the spatiotemporal region of the coin toss is *exactly* the way it is in *w*; not so in an antecedent-world in which the coin toss has a different outcome. Hence, the antecedent-world in which the coin toss has the same result has a greater spatiotemporal region of perfect match. Lewis's view therefore yields the correct result that (6) is true.

signal passes out of it, except for the one travelling to the display. Now consider the following counterfactual:

(7) If the random process in *A* had turned out differently and the signal had travelled from *A* to *C* rather than to *B*, ticket number 17 would still have won.

If the signal had travelled from *A* to *C*, then the outcome of the draw would have been determined by a random process inside *C* rather than by one inside *B*. I take it that there is no reason whatsoever for thinking that the same ticket would have won in this case, and almost everyone I have asked about the case agrees with me. I therefore conclude that (7) is not true at *w*.[12]

But note that the closest antecedent-worlds in which the same number is drawn (and the result is transmitted to the display at exactly the same time, and by a signal of precisely the same kind) as in *w* is exactly like *w* throughout the post-antecedent time, with the only exception of the interior of the box. Nonetheless, such a world does not count as closer to *w* than another antecedent-world in which a different number is drawn and which differs from *w* throughout a massive part of post-antecedent space-time. This shows that, contrary to Lewis's theory, large regions of perfect match in matters of particular fact sometimes contribute nothing to closeness.

Now let us add another feature to the example. You are watching the lottery draw on television in your room hundreds of miles away. Just before the draw someone offers you to sell you ticket number 17, but you decline. Consider

(8) If you had bought the ticket, you would have won.

I think that this counterfactual is true in *w*. And so do most people I have asked. But the truth of (8) presupposes that ticket number 17 would still have won if you had bought that ticket. It therefore seems that in this example the similarity of an

[12] A very similar example was developed simultaneously and independently by Wasserman (see his (forthcoming)) to make essentially the same point.

antecedent-world with respect to the outcome of the lottery draw does contribute to closeness.

Why do we think that the same ticket would have won if you had bought the ticket, but not that the same ticket would have won if the signal had travelled from *A* to *C* rather than to *B*? What is the relevant difference between the two cases? I think that by far the most natural thing to say is something along the following lines: There is no causal connection between your decision about buying the ticket and the outcome of the lottery draw. In a world in which you purchase the ticket, the causal history of the outcome is just the same as in *w*. By contrast, in a world in which the signal travels from *A* to *C* rather than to *B* and the winning ticket is determined by a random process inside *C*, the causal history of the outcome is different.

This suggests the following general principle:

> (c) If a matter of particular fact obtains in two worlds, then this contributes to the closeness between the two worlds if and only if the relevant matter of fact has the same causal history in the two worlds.[13]

The formulation of this principle is provisional only and will be generalized and revised in the next section. For now we should merely note the implications for Lewis's account of the results surveyed in this section. We have seen that the phenomena that illustrate (c) present an extensional problem for Lewis's account. If we wanted to revise his theory so as to solve this problem, we would need to modify his clauses relating to perfect and approximate match in such a way as to accommodate the findings of the foregoing discussion. This would be bad enough. It would be the second complication (after the clause about quasi-miracles) that needs to be added to the four-clause account that Lewis propounds in the passage quoted in section 2. But I think that the findings of this section give rise to an objection that cuts

[13] I intend to leave it open which kinds of entities (facts, events, states, etc.) can count as 'matters of particular fact' in the sense relevant to (c). I will sometimes write as if they included only facts, and as if only facts could figure in the causal history of other facts. However, I do so merely for the sake of convenience. I do not mean to commit myself to an account of causation according to which the *relata* of the causal relation are facts. I think that anything I say could be reformulated in a way that makes it clear that it incurs no commitment to any specific view about what the *relata* of causation are.

deeper than considerations of extensional adequacy: I think that they undermine the *motivation* for some of the central principles of Lewis's account. Remember that Lewis's sole reason for maintaining that similarities in regions of perfect match contribute more weightily to closeness than those in regions of approximate match is the wish to explain the asymmetry of counterfactual dependence. The same is true for the assumption that quasi-miracles detract from closeness. Now, even if we accepted these two assumptions, we would need to somehow incorporate the result of this section—some principle like (c)—into the account of the closeness relation. But, as I will try to explain in more detail in section 7, principle (c) (or, rather, the generalized version of it to be propounded in the next section) alone can do all the work in the explanation of the asymmetry of counterfactual dependence that Lewis's two assumptions were intended to do. Hence, once we decide to incorporate a principle like (c) into our account of the closeness relation, there is no work left to do for Lewis's two assumptions. There is no reason for retaining such idle baggage.

4. Closeness and explanation

In section 1, we considered two principles—(2) and (3)—about the standard truth-conditions of counterfactuals: Firstly, counterfactual dependence is temporally asymmetrical. Secondly, the conformity of an antecedent-world to the laws of the actual world contributes to its closeness. As we saw in section 2, Lewis noted that the asymmetry of counterfactual dependence was not written into the truth-conditions of counterfactuals, but that it ought to be explained by appealing to certain features of our world. But he did not take a similar view about (3), the principle that the conformity of an antecedent-world to the actual laws contributes to its closeness. According to the four-clause formulation of his account which I quoted in section 2, it is written into the truth-conditions of counterfactuals (without qualification or restriction to specific kinds of counterfactuals) that nothing detracts from the closeness of an antecedent-world as much as a big miracle does, and that small miracles, too, detract from closeness. Hence, on Lewis's account, it is true for every

counterfactual whatsoever, and in any possible world *w*, that the conformity of an antecedent-world to the laws of *w* contributes to its closeness to *w*.

This view is to be strictly distinguished from the thesis that, no matter which counterfactual we are considering, all the actual laws hold in the closest antecedent-worlds. The latter view is hardly an option. Many counterfactuals whose antecedents are inconsistent with the actual laws seem perfectly intelligible, for example 'If Fred's Toyota were faster than light, he could drive to New York in under one minute.' In such cases, the closest antecedent-worlds are presumably worlds in which some actual laws fail to hold. However, although it cannot be true of all counterfactuals that all actual laws hold in all the closest antecedent-worlds, it might nonetheless be true for all counterfactuals that the conformity of an antecedent-world to the actual laws contributes to its closeness. All worlds in which Fred's Toyota moves faster than light contain some violations of the actual laws, but some conform to the actual laws more closely than others. Some contain no counterlegal events except those connected with the Toyota's exceptional performance, others are alive with the most blatant and appalling violations of actual law. It seems open to a philosopher to hold that, all other things being equal, worlds of the first kind are closer than those of the second.

It is thus not obviously absurd to maintain, with Lewis, that (3) is true of all counterfactuals whatsoever, including those whose antecedents contradict the actual laws. However, I think that this view is not true. I will argue that (3) holds for some counterfactuals but not for others. I believe that in the case of some counterfactuals (including some whose antecedents are consistent with all actual laws), an antecedent-world's degree of conformity to the actual laws is simply irrelevant to its closeness.

My argument for this conclusion will rest on a presupposition that is perhaps more controversial than the assumptions underlying my discussion in the previous sections. This presupposition could be stated as follows:

(L) Where *L* is a law of nature and *E* is a course of events that instantiates *L*, the fact that *L* is a law is one of the factors that jointly explain *E*.

> Moreover, for any law L, the fact that L is a law explains the general fact that matters of particular fact conform to L (i.e. the fact that L is a law explains why L is true).[14]

To take an example, consider

> (Law of Gravitation) Any two bodies of masses m_1 and m_2 that are at distance d of each other attract one another with a force of strength Gm_1m_2 / d^2,

where G is the gravitational constant. Assume that (Law of Gravitation) is a law of nature. During last week, Mars took a certain path through space in accordance with (Law of Gravitation). I believe that the fact that (Law of Gravitation) is a law is one of the factors that jointly explain why Mars took the path it did. And I also believe that the fact that (Law of Gravitation) is a law explains the general fact that events conform to (Law of Gravitation); that is, it explains why bodies of masses m_1 and m_2 that are at distance d of each other attract one another with a force of strength Gm_1m_2 / d^2.

(L) seems very plausible to me. Most other people I have asked find the principle plausible, too, and this makes me hope that the reader will find it reasonable as well. Unfortunately, any serious discussion of (L) would have to take up a lot of space, and is therefore beyond the scope of this paper.[15]

(L) will play a twofold role in the discussion of this section: *Firstly*, although my informal polls suggest that most people are happy to accept the intuitive judgments about individual counterfactuals on which the argument of this section rests, I suspect that a reader who does not accept (L) might disagree with these judgments. Since the

[14] In Sct. 5, I will suggest that laws can have exceptions, in the sense that a principle L can be a law in a possible world *w* even if there are exceptions to L in *w*. Once we accept such a view, we do not want to rule out the possibility that there are exceptions to the actual laws in our world. That is, we want to leave open the possibility that our world does not perfectly conform to all the actual laws. But even if our world only approximately conforms to a given actual law L, I think that this approximate conformity can still be explained by the fact that L is a law.

[15] Thanks to Harold Hodes, Marc Lange, Geoffrey Sayre-McCord and Thomas Hofweber for useful discussion of the point.

intuitive judgments are shared by so many, a reader who does not endorse them might still take an interest in the fundamental question of this section, viz. how these judgments are best to be accommodated by a theory of counterfactuals. *Secondly*, I will support the views propounded in this section by arguing that they offer an attractive explanation of the intuitive judgments at issue. This explanation, however, will be based on (L).

Suppose that (Law of Gravitation) is a fundamental law of our world, that is, that it is a law, and that this fact cannot be explained by appeal to other, more fundamental laws. Consider,

(9) If (Law of Gravitation) had not been a law, then events would still have at least approximately conformed to it.

No one I asked believed that this counterfactual was true.

The example seems to show that for some counterfactuals the range of antecedent-worlds that can count as the closest is not constrained by their degree of conformity to all the actual laws. Antecedent-worlds that approximately conform to (Law of Gravitation) do not for this reason count as closer than those that do not. This is so, despite the fact that the antecedent of (9) appears to be consistent with all laws, including (Law Gravitation). (It seems plausible that (Law of Gravitation) could have failed to be a law while still being *true*.[16])

Our reluctance to accept (9) presents a problem for Lewis's theory. Lewis assumes that big miracles detract more from closeness than anything else, so that big-miracle worlds can be among the closest antecedent-worlds only if there are no antecedent-worlds without big miracles. But in the case of (9) there are, of course, worlds of the

[16] The assumption that the antecedent of (9) is consistent with (Law of Gravitation) seems to be in line even with sophisticated regularity accounts of lawhood, such as Lewis's. On the account Lewis presents in 1973, p. 73, the laws are (roughly) those propositions that are theorems in every deductive system that provides one of the best trade-offs between the *desiderata* of simplicity and deductive strength. In our world (let us assume) (Law of Gravitation) is both true and a theorem of all the best systems. But presumably (Law of Gravitation) is also true in some possible worlds in which it is not a theorem of the best deductive systems. Consider worlds that contain no bodies with mass at all. (Law of Gravitation) is vacuously true in such worlds. But the axioms of the best deductive systems of such a world may not mention mass at all, and may therefore not yield (Law of Gravitation) as a theorem.

latter sort. (Any world that contains one small violation of (Law of Gravitation) and no other miracles, for example, is an antecedent-world without big miracles.) We can conclude that, on Lewis's account, none of the closest antecedent-worlds contain big miracles. The closest antecedent-worlds contain at most small and inconspicuous miracles. Hence, if (Law of Gravitation) had not been a law, there would still have been at most small and inconspicuous violations of it. Thus, on Lewis's theory we should expect (9) to be true. But that is contrary to intuition.

Let us consider another example that puts similar pressure on Lewis's theory. Suppose that there is one simple fundamental deterministic law, the 'master law.' All events conform to this law, and the fact that they do can be explained by the fact that the master law is a law. The lawhood of the master law explains the lawhood of all other laws. Assume further that time has no beginning, and that all matters of particular fact can be explained by appeal to the fact that the master law is a law and earlier matters of particular fact. Consider

(10) If the master law had not been a law, the history of the world would still have been very similar to what it was actually like.

I think that (10) is in the same position as (9): It does not intuitively seem to be true. Yet on Lewis's theory we should expect (10) to be true. The argument for this claim is a little more complex than in the case of (9). Let us go through it step by step. Assume that *w* is one of the closest antecedent-worlds. Either *w* contains miracles, or it does not.[17] Let us consider these two cases separately:

Case 1: Miracles occur in w. Among the miracle-worlds in which the antecedent is true, there are some that contain tiny and inconspicuous miracles, but no big ones. (Any world that contains a small and inconspicuous violation of the master law and no other miracles is of this kind.) On Lewis's account, such worlds are closer than antecedent-worlds with big miracles. Hence, no big-miracle worlds can be among the

[17] I do not think that every possible antecedent-world must contain a miracle. It seems to me that it is possible for the master law to be *true* without being a *law*. Hence, there can be possible worlds in which there are no violations of the master law, but the master law nonetheless fails to be a law.

closest antecedent-worlds. *w* therefore contains only small and inconspicuous miracles.

According to Lewis's theory, a small-miracle antecedent-world will be the closer the greater its region of perfect match. We therefore presumably have to say that *w*, being one of closest small-miracle antecedent-worlds, is exactly like our world before the miracle and with respect to all space-time regions not affected by the miraculous events. This means that there is a massive amount of match in matters of particular fact between *w* and our world. The consequent of (10) is therefore true in *w*.

Case 2: w contains no miracles. Lewis assumes that small and inconspicuous miracles can be outweighed by increases in the size of the region of perfect match. This means that a miracle-free antecedent-world can be among the closest antecedent-worlds only if no antecedent-worlds with tiny miracles have greater regions of perfect match. We already know that there are small-miracle antecedent-worlds with massive regions of perfect match. Hence, given that *w* is, by hypothesis, both miracle-free and among the closest antecedent-worlds, *w* too must contain large regions of perfect match. The consequent of (10) must therefore be true in *w*.

On Lewis's theory the consequent of (10) is true in all the closest antecedent-worlds, no matter whether they contain miracles or not. Lewis's account therefore yields the intuitively problematic consequence that (10) is true.

How can we explain our unwillingness to accept conditionals like (9) and (10)? I will focus on an explanation that, I think, is supported by the fact that it has considerable intuitive force. Consider (9) again. The only reason why events conform to (Law of Gravitation) in our world is that (Law of Gravitation) is a law. But that reason is absent in an antecedent-world. Hence, even if the events of an antecedent-world (approximately or even perfectly) conform to (Law of Gravitation), their conformity to the law does not have the same explanation as in our world and the world's similarity to our world with respect to this conformity therefore contributes nothing to closeness. Antecedent-worlds that conform to (Law of Gravitation) are no closer than those that do not. This is why there is no reason for accepting (9).

Consider (10) next: One part of the reason why the history of our world unfolded in the way it did is that the master law is a law. But the master law is not a law in any

antecedent-world. Hence, even if the history of an antecedent-world is similar to that of our world, the reason why history unfolds that way in it (if there is any reason) is not the same as in our world. The similarity in the histories of the two worlds therefore contributes nothing to closeness.

The foregoing considerations suggest the following principle:

> (C) If some fact *f* obtains in both of two worlds, then this similarity contributes to the closeness between the two worlds if and only if *f* has the same explanation in the two worlds. (In the special case in which *f* has no explanation in either world, this condition counts as vacuously satisfied.)[18]

As we will see in section 8, this formulation is a little too simple. But I think that it can serve as a working account for our present purposes.

(C) is compatible with different ways of understanding the term 'explanation,' including some on which it expresses an epistemic notion. But I prefer to combine principle (C) with a version of what is sometimes called an 'ontic' conception of explanation (Salmon 1984). On my interpretation, the term 'explanation' expresses an objective metaphysical, non-epistemic relation: To say that the fact *f* is one of the facts that jointly explain the fact *g* is to say that *f* is part of the reason why *g* obtains, that *f* is one of the factors that jointly gave rise to, or are responsible for, *g*. I take causation to provide the paradigmatic examples of the relation I have in mind: If *X* is a cause of *Y*, then *X* is one of the factors that are jointly responsible for, or explain, *Y*. (The reason why the ball started to move is that the player kicked it. The kick is one of the factors that explain the ball's beginning to move.) However, I think that the relation of one thing's explaining another is more general than that of causation, that it can hold between things that cannot cause each other. We have already considered

[18] It may be objected that there are similarities in facts with the same explanation that do not contribute to closeness, for example similarities with respect to disjunctive facts. (Perhaps it contributes nothing to the closeness between two worlds that emeralds are grue (i.e. either green and first observed before 2000, or blue and first observed after 2000) in both of them, even if the explanation of this fact is the same in the two worlds.) In order to accommodate this point, one might have to impose some restriction on the range of facts that (C) quantifies over. There are several ways in which this might be done. Unfortunately, I have no space to discuss them.

an example: I believe that the fact that (Law of Gravitation) is a law explains why events conform to this law (but it would be odd to say that the lawhood of the law *causes* events to conform to the law). Similarly, I think that the lawhood of one law can explain the lawhood of another; for example, the fact that (Law of Gravitation) is a law might explain why Kepler's Laws are laws. I will also assume that the relation of explanation is transitive. If *f* is one of the factors that explain *g* and *g* is one of the factors that explain *h*, then *f* is one of the factors that explain *h*.

I suggested that causation is a special case of one fact's contributing to explaining another, that is, that the concept of one fact's explaining another is a generalization of the notion of causation. According to principle (c), the sharing of a matter of particular fact *f* between the actual world and an antecedent-world contributes to the closeness between the worlds only if *f* has the same causal history in the two worlds. Principle (C) imposes a generalization of this requirement: In order for the sharing of any fact *f* between the actual world and an antecedent-world to contribute to closeness, the explanation of *f* must be the same in the two worlds. In this sense, the result of the discussion of this section is a generalization of that of section 3.

I think that the assumption that some principle like (C) is true can be strongly supported by its explanatory power. As a generalization of (c), (C) can explain our intuitions about the sorts of examples that support (c), for example the three-random-generator case of section 3. It can also explain our pre-theoretical judgments about examples like (9) and (10) that cannot be explained by (c). Moreover, in section 7, I will argue that (C) can be used to explain the asymmetry of counterfactual dependence. Finally, as I will explain in section 9, I believe that the explanatory power of (C) extends even further than that. Given the explanatory significance of the principle, I think it deserves a central role in an account of standard counterfactuals.

5. The laws and the asymmetry of counterfactual dependence

Analytic philosophers often (though by no means universally) assume that a universal generalization cannot be a law of nature unless it is true without exception (i.e. that a universal generalization L cannot be a law in a possible world *w* unless L is true

without exception in *w*). Let us assume for the moment that this is correct. Now, we noted in section 1 that, if the antecedent of a counterfactual is a falsehood about matters of particular fact, and if the world is deterministic (in the sense defined in section 1), then any possible antecedent-world must either contain some violation of the actual laws, or it must be unlike our world throughout the pre-antecedent time. If laws cannot have exceptions, then a possible antecedent-world that contains a violation of the actual laws must be one in which different laws are in force. So, it seems that under determinism, we need to adopt at least one of the following two counterfactuals (their antecedent is false):

(11) If I had not scratched my nose a minute ago, the history of the world might have been different throughout the period before the scratching.

(12) If I had not scratched my nose a minute ago, the laws of nature might have been different.

Both counterfactuals seem pretty implausible to me, so it seems that we have arrived at a dilemma.

I think that the dilemma can be made to seem even more repugnant than it might appear at first blush. Our discussion in sections 3–4 should already make us suspect that there is a substantial problem. Assume first that we adopt (12), and say that among the closest antecedent-worlds there is some world *w* in which some actual law L fails to be a law. Now, the lawhood of L presumably figures in the actual explanations of many matters of particular fact of our world. Hence, even if these matters of fact obtain in *w*, they cannot have the same explanations as in our world, so that, according to principle (C), the fact that they obtain in *w* contributes nothing to the closeness of *w*. Hence, there might be other antecedent-worlds in which the relevant matters of fact do not obtain, but which have as many *closeness-relevant* similarities to our world as *w*. Since *w* is among the closest antecedent-worlds, these other antecedent-worlds must be among the closest as well. We arrive at the

conclusion that the closest antecedent-worlds may include some that are implausibly dissimilar from our world in matters of particular fact.

Similar problems arise if we seize the other horn of the dilemma by adopting (11). In this case, we need to say that among the closest antecedent-worlds there is some world *w* in which many of the matters of particular fact that obtain in our world during the pre-antecedent time fail to obtain. But these pre-antecedent matters may figure in the actual explanations of countless other matters of particular fact before, at, and after the antecedent-time. Hence, even if these other matters of fact obtain in *w*, they do not have the same explanations in *w* as in our world, and are therefore, according to (C), irrelevant to the closeness of *w*. Hence, there might be other antecedent-worlds in which the relevant matters of fact do not obtain, but which have as many closeness-relevant similarities to our world as *w*, so that these worlds, too, must be among the closest antecedent-worlds. Once again, we arrive at the conclusion that the closest antecedent-worlds may include some that are implausibly dissimilar from our world.

In generating this problematic consequence, I have made explicit use of principle (C). The reader may therefore think that it is my principle (C) that burdens us with this difficulty, and that the result counts strongly against (C). I think, however, that principle (C) is not needed to generate the problem. Instead of using (C), we can directly appeal to the intuitions that motivated principle (C) in the first place. The problem is therefore bound to arise as long as we accept those deliverances of intuition that motivate (C), whether or not we use (C) to explain these intuitions.

Let me choose a particularly dramatic example to illustrate this point. Assume that in our world time has a first moment, and that at that moment all matter was concentrated in a single small sphere. Suppose that the entire remaining course of the history of the world can be deduced from the density of mass in that sphere at the first moment and the laws of nature. The sphere's density of mass, which from now on I will simply call the 'master parameter,' contributes to explaining all subsequent matters of particular fact. The course of history is extremely sensitive to small variations in the value of the master parameter. Possible worlds in which the density of the sphere is just minimally different and which evolve in accordance with the

actual laws often have strikingly different histories. Assume further that there is also a master law with the properties already described: The lawhood of the master law explains the lawhood of all other laws and contributes to explaining all matters of particular fact of our world.

If laws cannot have exceptions, then we need to say that at least one of the following conditionals is true:

(13) If I had not scratched my nose a minute ago, then the master parameter might not have had the value it did.

(14) If I had not scratched my nose a minute ago, then the master law might not have been a law.

It can be argued that adopting either of these two conditionals will lead to a problem. My attempt to show this will proceed in three steps:

i. Since practically all matters of particular fact of our world obtain only because the master law is a law, it seems intuitively that there is no reason for thinking that the history of the world would have taken a similar course if the master law had not been a law. Rather,

(15) If the master law had not been a law, then the entire history of the world might have been completely different.

In particular, if the master law had not been a law, the history of the world might have taken a completely different course and I might never have been born. Now, consider the question:

What would have happened if I had not scratched my nose a minute ago and the master law had not been a law?

We have already seen that the closest worlds in which the master law is not a law include some whose entire history is completely different from that of our world and in which I am never born (and therefore never scratch my nose). These worlds, since they are among the closest of those in which the master law fails to be a law, must a fortiori be among the closest of those in which the master law fails to be a law and I do not scratch my nose a minute ago. Hence, we need to conclude that

(15*) If the master law had not been a law and I had not scratched my nose a minute ago, then the entire history of the world might have been completely different.

ii. The matters of particular fact of our world obtain only because the master parameter had the value it did. Even minimally different values would have led to strikingly different histories. It therefore seems intuitively that there is no reason for thinking that the matters of particular fact of our world would still have obtained if the master parameter had not had the value it did. Rather, intuition tells us that

(16) If the master parameter had not had its actual value, then the entire history of the world might have been completely different.

But once we accept (16), then reasoning exactly parallel to that which led us from (15) to (15*) will lead us to accept

(16*) If the master parameter had not had its actual value and I had not scratched my nose a minute ago, then the entire history of the world might have been completely different.

iii. Now suppose that (13) is true. Then the closest worlds in which I did not scratch my nose a minute ago include some in which the master parameter has a different value. But, since (16*) is true, the closest worlds in which I did not scratch my nose a minute ago and the master parameter has a different value must include some whose history is completely different from that of the actual world. We can conclude that the closest worlds in which I did not scratch my nose include some whose history is entirely different from that of our world. We thus obtain the consequence that

(17) If I had not scratched my nose a minute ago, the entire history of the world might have been completely different.

Thus, if (13) is true, then (17) must be true as well. We can use exactly analogous reasoning (appealing this time to (15*) rather than to (16*)) to show that (17) must be true if (14) is true. Hence, if either (13) or (14) is true, then (17) must be true.

The assumption that at least one of the two conditionals (13) and (14) is true in our example therefore yields the consequence that (17) is true in the example as well. (Note that I have not used (C) in this argument, but have instead directly appealed to the kind of intuition that motivates (C), viz. the intuition that (15) and (16) are true in our example.) Now, accepting that (17) is true in the example seems to be an unattractive option. After all, for all we non-scientists know, it might be that the situation envisaged—the scenario in which there is a master law and a master parameter that explain everything—obtains in the *actual* world. So, if we accept that (17) is true in the example, how can we be sure that (17) is not *actually* true?

One possible response to this problem is to deny one of the intuitive assumptions on which the argument rests, namely either the assumption that (15) is true in the example envisaged, or the assumption that (16) is true in the example. Of course, since principle (C) underwrites these two assumptions, if we wanted to give up either of them, then principle (C) would need to be weakened in some suitable way so as no longer to commit us to the relevant assumptions.

There is, however, another response to the problem that I prefer to the one just considered, and that is to give up the assumption that generated the dilemma in the first place, namely the assumption that laws cannot have exceptions. It seems to me that, inasmuch as we are in the business of trying to capture the pre-theoretical notion of lawhood, this assumption ought to be controversial. That is, it does not seem obvious to me that there is anything in the folk concept of a law that precludes the existence of exceptions to a law. For many centuries belief in miracles was very common as a central component of popular religious faith, and on one natural and common way of understanding the notion of a miracle, it involves a violation of natural law.

Suppose that we allow that laws can have exceptions, that is, that it can be a law that all *F*'s are *G*, even if there is a small number of exceptions to this universal generalization. Once we do that, we can maintain that there are possible worlds which contain small violations of the actual laws, but which nonetheless have the same laws as our world. This allows us to say that, even under determinism, there are antecedent-worlds that are like ours until shortly before the antecedent-time and which have the same laws (even though they contain small violations of these laws), and that neither (11) nor (12) is true. (This way of avoiding the choice between (11) and (12) was first suggested to me by Brian Weatherson. Marc Lange has defended a similar line in detail in his (2000).) Such a view seems to me to be worth exploring, and this is what I will do in the rest of this paper.

I think that much more could be said to motivate the assumption that laws of nature can have exceptions (and much more *has* been said, e.g. by Marc Lange in his (2000)). Unfortunately, I cannot discuss the matter in detail here. Needless to say, I realize that some readers might be unwilling to accept that laws can have exceptions, and such readers will not accept my resolution of the difficulty presently under consideration. Such readers are referred to the appendix, in which I present one of the several possible alternative solutions to the problem that are consistent with the assumption that laws cannot have exceptions.

6. Nomic necessities

Without attempting to conform precisely to previous philosophical usage, I will call all truths about which principles are laws of nature (true propositions of the form *P is a law* and *P is not a law*) and all the propositions that follow from these truths 'nomic necessities.' Propositions that are compossible with all nomic necessities, and possible worlds in which all nomic necessities hold, will be called 'nomically possible.' (A possible world is nomically possible if and only if it has the same laws as our world (whether or not it perfectly conforms to these laws).) It seems plausible to me that when we reason about what would have been the case if a certain nomically possible antecedent had been true, we assume that events would more or less still have conformed to all actual laws.[19] When reasoning from nomically possible falsehoods about matters of particular local fact, for example, we assume that the course of events initiated by the antecedent-event would have unfolded in perfect accordance with all the actual laws. (This why we want to say that (1) is true if determinism is true and the missile system is in flawless working order.) It thus seems to be true for any counterfactual whose antecedent is nomically possible that the degree to which an antecedent-world conforms to any given actual law is relevant to its closeness. According to my account, this means that the conformity to each of the actual laws must have the same explanation as in our world. Since (according to principle (L) of section 4) in our world it is the lawhood of the relevant laws that explains why events conform to them,[20] this must also be true in the closest antecedent-worlds. But this means that all the actual laws must be laws in the closest antecedent-worlds as well.

On the other hand, when reasoning about what would have been the case if some nomically possible antecedent had been true, we do not assume that any laws that are not in force in our world would have been in force and would have placed constraints on the causal chain initiated by the antecedent-event. We assume that the causal chain initiated by the antecedent-event might have taken *any* course that is compossible

[19] I believe, following Lewis, that small miracles occur in the closest antecedent-worlds. But I also think that miracles are the exception and that the closest antecedent-worlds display at least approximate conformity to each of the actual laws.

[20] More precisely: In our world, the lawhood of a law L contributes to explaining why events conform to L to the degree that they do. Of course, it could be that the conformity of the actual events to L is only approximate. Even in this case, the fact that events approximately conform to L is explained by the fact that L is a law.

with the actual laws and with those matters of particular fact that we hold fixed. *Any* such way for the causal chain to unfold is realized in some of the closest antecedent-worlds.[21] (Suppose that there is some chance of a nuclear explosion and some chance that the signal will fizzle out when Nixon's button is pressed. In this case, we want to say that there might have been a nuclear explosion or the signal might have fizzled out if Nixon had pressed the button. Each of the two lawful courses of events is realized in some of the closest antecedent-worlds.) This indicates that the closest antecedent-worlds do not have any laws except those that are also laws in the actual world.

These considerations suggest that,

(18) for any nomically possible antecedent, the closest antecedent-worlds are nomically possible worlds, that is, metaphysically possible worlds with the same laws as ours.

(18) entails that any world with the same laws as ours is closer than any world with different laws.[22] We can conclude that sameness of the laws contributes more weightily to the closeness between worlds than any similarities in particular fact, and therefore outweighs any dissimilarity in particular fact.

---

[21] In one special case, Lewis would disagree with the claim that the chain of events initiated by the antecedent-event might have taken any lawful course: A quasi-miracle is perfectly lawful, and yet (as we saw in footnote 6), Lewis is committed to saying (implausibly in my view) that there would have been no quasi-miracle if Nixon had pressed the button.

[22] In proving that (18) entails that the nomically possible worlds are closer than the nomically impossible worlds, I will make use of the following principle:

(28) For any possible world *w*, there is some proposition that is true in *w* and in no other possible world.

Now suppose that (18) is true and assume for *reductio* that it is not true that the nomically possible worlds are closer than the nomically impossible ones. Then there must be some nomically impossible world w_i and some nomically possible world w_p, such that w_i is at least as close to our world as w_p. According to (28), there is some proposition P_{wi} that is true in w_i and in no other possible world, and a proposition P_{wp} that is true in w_p and in no other possible world. Now consider the proposition P_{wi} *or* P_{wp}. This proposition is true in the nomically possible world w_p and must therefore itself be nomically possible. Hence, according to (18), the closest possible P_{wi}-*or*-P_{wp}-worlds must be nomically possible. Since w_p and w_i are the only possible P_{wi}-*or*-P_{wp}-worlds, and since w_p is nomically possible while w_i is nomically impossible, we can conclude that w_p must be closer than w_i. But this contradicts the assumption that w_i is at least as close as w_p.

7. Explaining the asymmetry of counterfactual dependence

Assume that we resolve the problem described in section 5 by allowing laws to have exceptions. Once we do so, I think that we can formulate a plausible account of counterfactuals—an account that, among other things, provides an attractive explanation of the temporal asymmetry of counterfactual dependence.

The explanation of the asymmetry that I will offer rests on certain suppositions about the closeness ordering as well as certain assumptions about the world:

Assumptions about the closeness ordering:

(19) (C)

(20) Nomically possible worlds (i.e. possible worlds with the same laws as ours) are closer than nomically impossible worlds.

Assumption (20) guarantees that, provided the antecedent is nomically possible, the same laws will be in force in all the closest antecedent-worlds. If laws could not have exceptions, then such a clause would guarantee, not only that the closest antecedent-worlds have the same laws as our world, but also that these laws are never violated in them. But if laws can have exceptions, then the fact that some actual law is also a law in the closest antecedent-worlds does not entail that this law is not violated in these worlds. (20) therefore places no very stringent constraints on the number, nature and spatio-temporal location of miracles in the closest antecedent-worlds. But presumably such constraints are needed. We do not want the closest worlds in which Nixon presses the button to feature, say, gratuitous miracles in the chain of events that is initiated by the button-pressing. Antecedent-worlds in which the signal miraculously disappears in the wire after the button-pressing are not among the closest, even if they have exactly the same laws as our world. Therefore, if we allow laws to have exceptions, then a principle like (20) that assigns great weight to sameness of laws needs to be supplemented by a rule to the effect that it detracts from the closeness of a world if it contains violations of the actual laws that do not occur in the actual world.

I suggest that we supplement (20) by two principles about conformity to the actual laws that are very similar to those propounded by David Lewis (see section 2). Call a miracle occurring in another world 'alien' just in case the same miracle does not occur in the actual world. Our two additional principles can then be stated as follows:

Assumptions about the closeness ordering (continued):

(21) A small and inconspicuous alien miracle detracts from the closeness of a nomically possible world, but can be outweighed by increases in similarity in matters of particular fact with the same complete explanations.

(22) A big alien miracle detracts far more from the closeness of a nomically possible world than small alien miracles do; so much so that big alien miracles cannot be counterbalanced by increases in similarity in matters of particular fact.[23]

Assumptions about the world:

(23) Causes precede their effects.

(24) The divergence of a world from ours is possible by a small miracle or without any miracle. (More precisely: For most of the nomically possible antecedents about matters of particular local fact that occur in ordinary-life counterfactuals, there is some possible antecedent-world that is exactly like ours until shortly before the antecedent-time and which contains at most a small and inconspicuous alien miracle.[24])

[23] As we saw in Sct. 4, the degree to which the events in another world *w* conform to the laws of the actual world is relevant to the closeness between the two worlds only if these laws are also laws in *w*. Hence, it is not true of *every* world *w* that the occurrence in *w* of an alien violation of some actual law L detracts from the closeness of *w*. (If L is not a law in *w*, then the miracle might be irrelevant.) However, since all actual laws are also laws in a nomically possible world, the occurrence of alien miracles does detract from the closeness of a *nomically possible* world.

[24] I do not claim that this is true of *all* counterfactuals whose antecedents are nomically possible and describe matters of particular local fact. Consider, for example, 'If Mount Everest were now twice as high as it actually is, ...' It might be that any antecedent-world that is just like ours until shortly before the antecedent-time and in which Mount Everest then quickly doubles in height contains a big alien miracle.

(23) might not hold universally. I think that exceptions to (23) are most likely also exceptions to our *explanandum* phenomenon, the asymmetry of counterfactual dependence. Let us therefore not worry about such cases.

To see how these assumptions do their work, suppose that Fred filed his tax return late this year and was penalized. Let us use as our example the conditional

> If Fred had filed his tax return two hours before the deadline, he would not have been penalized.

Let us first consider the case in which the world is deterministic (in the sense defined in section 1[25]). We can appeal to (24) to justify the claim that there is some possible antecedent-world that is just like our world until some time *t* shortly before the antecedent-time, then smoothly diverges by a small alien miracle, and after that unfolds without any alien miracles. Let *w* be some world of this kind and assume that the laws of *w* are the same as those of our world (i.e. that *w* is nomically possible). (Given the assumption that laws can have exceptions, this is consistent with the assumption that some actual laws are violated in *w*.) Assuming that the IRS is perfectly reliable in assigning late-filing penalties, *w* must be a world in which Fred does not get a penalty. The above assumptions guarantee that worlds like *w* are the closest antecedent-worlds.[26]

---

[25] Remember that on the definition stated in Sct. 1, determinism is the thesis that any two worlds that perfectly conform to the laws of our world and are alike throughout some extended initial segment of their histories are alike throughout their histories. (Note that, on the present assumption that laws can have exceptions, the actual world might not be among the worlds that perfectly conform to the actual laws.)

If we allow for the possibility that laws have exceptions, then we cannot instead formulate determinism as the thesis that any two possible worlds that are perfectly alike throughout some extended initial segment of their histories *and which have the same laws as our world* must be alike throughout their histories. If laws can have exceptions, then this thesis can never be true. For a world can have the same laws as our world and yet not perfectly conform to these laws. Hence, no matter what the actual laws are like, there will be two worlds that have the same laws as our world and which are alike throughout some extended initial segment of history that ends at time *t*, and one of which features a violation of law after *t* while the other one does not, so that the two worlds are not entirely alike throughout their histories.

[26] More precisely: The above assumptions guarantee that worlds like *w* are the closest antecedent-worlds, except in certain special cases. Suppose that in our world there was some time *t* shortly before Fred's deadline for submitting his tax return, such that the history of the world up to *t*, together with

In order to see this, note first that, given the temporal asymmetry of causation, the only matters of fact that figure in the causal histories of matters of fact before *t* in the actual world are themselves located before *t*. Hence, given the exact match between the actual world and *w* before *t*, we can assume that the matters of fact before *t* have the same causal histories in the two worlds. And since the lawhood of the same laws figures in their explanations, we may assume that the matters of fact before *t* have the same explanations in the two worlds. The pre-divergence similarities between *w* and our world therefore matter to closeness.

We can use this observation in explaining why antecedent-worlds like *w* are closer than any others. Given assumption (20), nomically impossible antecedent-worlds are not even candidates for the title of closest antecedent-world. So, what needs to be shown is that worlds like *w* are the closest among the nomically possible antecedent-worlds. Among the nomically possible antecedent-worlds, there are three types of worlds that are rivals of *w* for the title of closest antecedent-world: Those with fewer alien miracles than *w*, those with greater pre-antecedent match in matters of particular fact, and those with greater post-antecedent match. Let us consider worlds of these three types, in each case showing that they are less close than *w*:

First, consider worlds like w_1: w_1 contains no alien miracles at all, and (given determinism) it therefore differs from our world before *t*.[27] But it is very much like our world around the antecedent-time, except that the antecedent is true in it. *w* contains a small alien miracle while w_1 does not. But *w* is more relevantly similar to our world in matters of particular pre-antecedent fact. Hence, by (21), *w* closer to the actual world than w_1.

Secondly, there are antecedent-worlds, exemplified by w_2, that are like our world *right until* the antecedent-time. In w_2, as in our world, Fred is drinking in a bar

---

the laws, determine that Fred does submit his tax return on time (in the sense that Fred submits his tax return on time in every possible world that is like our world until *t* and which perfectly conforms to the actual laws thereafter), but that some law is violated in our world after *t*, so that Fred does not submit his tax return on time after all. In this case, the closest antecedent-worlds might be worlds that are like our world until *t* and which simply omit the fateful miracle that led to Fred's filing his tax return late. Such antecedent-worlds may not contain any alien miracles, and may in this regard be unlike *w*.

[27] As already mentioned in the preceding footnote, even under determinism there are special cases in which there are possible antecedent-worlds that contain no alien miracles but are nonetheless like our world until *t*. As I said in the preceding footnote, these cases are exceptions to the general principle that under determinism worlds like *w* are the closest antecedent-worlds.

immediately before the antecedent-time, not having prepared his tax return. At the antecedent-time Fred suddenly and miraculously disappears from the bar and reappears in the post office, sending off a completed tax return form that has just spontaneously popped into existence. As concerns matters of particular pre-antecedent fact, w_2 is more similar to our world than w. But in contrast to w, w_2 contains a big and conspicuous alien miracle. According to (22), w_2 is therefore less close.

Thirdly, there are worlds like w_3: w_3 smoothly diverges from our world by a small alien miracle shortly before the antecedent-time so as to make the antecedent true. After the antecedent-time, a tiny alien miracle happens in one of the computers at the IRS, and Fred is given an unjustified late-filing penalty. w_3 is more similar to our world than w with regard to some post-antecedent matters, viz. the penalty and events that happen as a consequence of it. But in w_3 the causal history of the penalty is different from what it is in our world. (In the actual world, the fact that Fred's letter to the IRS carried a postage stamp with a late date on it figures in the causal history of the fine, in w_3 it does not.) According to (C), the similarity between w_3 and our world with respect to the fine is therefore irrelevant to closeness. During the post-antecedent time w_3 is no more *relevantly* similar to our world than w is. Since w_3 contains two small alien miracles while w contains only one, w is closer.

So much about the deterministic case. Consider the case of indeterminism next. We can show by reasoning analogous to the one we have just gone through that the closest antecedent-worlds are like ours until shortly before the antecedent-time, and that they contain no alien miracles after that, even if such miracles would allow for approximate reconvergence to our world. However, there is an additional complication to consider. Under indeterminism (approximate and even perfect) reconvergence might be possible *without* any miracle. Suppose that the IRS is not entirely reliable, so that there is some chance that a timely filer will get a penalty. In that case, there are two kinds of antecedent-world to consider. On the one hand, there are worlds like w_4, in which the IRS does its job properly and Fred is not fined. On the other hand, there are worlds like w_5 in which some unfortunate low-chance event happens in the brain of some IRS employee and Fred gets an unjustified penalty. It

seems to me that, if there is some chance for a timely filer to get a penalty, then we should not say that Fred would *not* have been penalized if he had submitted his tax return on time. Nor, of course, should we say that he *would* have been penalized. He might or might not have. In other words, we want the result that worlds like w_4 and those like w_5 are equally close.

Assumption (C) secures this result: In worlds like w_5 the causal history of the fine is different from what it is in our world. Hence, although worlds like w_5 are more similar during the post-antecedent time than worlds like w_4, these extra similarities count for nothing. The former worlds are no closer than the latter.

I mentioned in section 2 that the right account of counterfactuals should be able to explain the apparent temporal asymmetry of counterfactual dependence, and that it should do so by appealing, not only to the truth-conditions of counterfactuals, but also to certain features of the world. The account I outlined clearly meets this criterion, since it entails that the temporal asymmetry of counterfactual dependence is ultimately grounded in that of causation (as well as in the feature of our world that is described by (24)).

Let us take stock. On the present account, the closeness of a world to a given world *w* is governed by the following system of weights:

1. It is of the first importance to ensure sameness of laws.
2. It is of the second importance to avoid big alien violations of the laws of *w*, provided the conformity to the relevant laws has the same explanation as in *w*.
3. It is of the third importance to maximize match in matters of particular fact with the same explanations as in *w*.
4. It is of the fourth importance to avoid small alien violations of the laws of *w*, provided the conformity to the relevant laws has the same explanation as in *w*.

Let me end this section by noting a difficulty for the view I presented. The problem concerns a type of counterexample of which I will present a specimen that is due to Pollock (Nute 1980, p. 104): I forgot my coat in the bar last night. In the course of the night two potential coat thieves passed by the coat, one at 10, the other at midnight. Each time, there was a non-zero chance that the coat would be stolen. The next morning I find to my relief that the coat is still where I left it. Now consider:

If the coat had been stolen last night, it would have been stolen at midnight.

There appears to be no reason for thinking that this counterfactual is true. We are inclined to think that, if the coat had been stolen, it might have been stolen at 10, or it might have been stolen at midnight. This example was originally presented as a counterexample to Lewis's view, but it applies to my account as well. I hold that, *ceteris paribus*, an antecedent-world is the closer the greater its match in matters of particular fact with the same explanations. The later an antecedent-world diverges from ours, the greater its match in such matters. My view therefore yields the implausible prediction that worlds in which the coat gets stolen at midnight are closer than those in which the theft occurs at 10.

The account I proposed in this section requires some revision in order to get around this problem. I have some hunches about possible solutions, but none that rise above the level of mere guesswork and conjecture. I will therefore merely note the existence of a difficulty. It remains a task for future work to offer a solution.

8. My account in need of refinement

The phenomena we have surveyed in sections 3 and 4 suggest that some principle very roughly like (C) is true. But, while (C) served us well as a working account of the phenomena we considered, I think that in its present formulation the principle is a little too simple.

As Cian Dorr and Kieran Setiya have pointed out to me, there are some examples that suggest that the left-to-right ('only if'-) direction of (C) is somewhat too strong. Suppose that we are watching an indeterministic lottery draw. The lottery was

instituted ten years ago. As part of this process, one of the people in charge, call her 'Susie,' made an important phone call. Susie has two qualitatively identical cell phones in her office, *A* and *B*. Every morning her secretary randomly chooses one of them and puts it on Susie's desk, and this cell phone will be used for all phone calls during that day. As it happens, on the day of the aforementioned phone call ten years ago, the secretary chose *A*, and this phone was used for the phone call. Consider:

(25) If Susie had used *B* rather than *A* for the aforementioned phone call ten years ago, the outcome of the lottery draw would have been just the same.

If the antecedent had been true, then the causal history of the lottery draw would have been different: Among the factors that figure in its *actual* causal history are certain processes inside *A*. By contrast, in the antecedent-world these processes are replaced by ones featuring the parts of *B*. And yet it seems to me that (25) is true, and this intuition seems to be shared by a large majority of those I asked in a little informal poll. (A small minority disagrees.)

Another difficulty for (C) also deserves attention. Note that we can distinguish between two different ways in which *x* can be part of the causal history of *y*. (The distinction I have in mind is inspired by a distinction that Ned Hall draws in his (2004), though it does not precisely coincide with the latter distinction.) On the one hand, *x* might part of what *produced y*. The stone throw produced the breaking of the window, the ingestion of poison produced the victim's death, and so forth. On the other hand, *x* it might be causally relevant to *y* without being among the producers of *y*. These non-producing factors include, for example, the absence of various kinds of possible interference with the causal process that produces *y*, as well as factors that prevent causal processes from interfering with the process that produces *y* and are therefore responsible for the absence of such interference. The absence of any assassins who try to kill the president on the eve of his speech figures in the causal history of the speech, as does the action of the police agent who arrests the assassin before he can strike. (It is partly *because of* the police agent and the absence of

assassins that the president holds his speech the next day). But neither the police agent nor the absence of assassins is among the factors that produce the speech.

The distinction between producing factors and other explainers applies to non-causal explainers as well as to causal ones. Where *L* is one of the laws in accordance with which the producing causes of *x* produced *x*, the fact that *L* is a law can be said to be among the factors that produced *x*. Now suppose that some causal process *P* prevented some other causal process *P*' from interfering with the process that produced *x* and was, in this way, partly responsible for *x*. The lawhood of the laws in accordance with which the elements of *P* prevented the interference from *P*' can in some sense be said to be among the factors that are jointly responsible for the obtaining of *x*, even though they need not be among the factors that produced *x*. When the king raises his hand and scratches his nose, there are countless laws of physics, biochemistry, etc. that are instantiated by the goings-on in the king's neurons, muscles, etc. The fact that these laws are laws is among the factors that are responsible for the nose-scratching, and can be said to be among the factors producing the scratching. The day before the scratching, an assassin who was about to kill the king the same evening was poisoned with cyanide by his enemies. The fact that it is a law that the ingestion of cyanide quickly leads to death is one of the factors that explain why the king was not murdered on the eve of the aforementioned nose-scratching, and is therefore one of the factors that explain why the nose-scratching took place. But it is not for this reason among the factors producing the nose-scratching.

We can thus distinguish between the *complete* explanatory history of a fact, that is, the totality of all factors that are explanatorily relevant to it, and the *productive* explanatory history, which consists only of those factors that produce the fact. This gives us two possible ways of understanding the idea of sameness of explanation as it occurs in (C). It could be interpreted as mere sameness of productive explanatory history, or it could be given the stronger reading as sameness of the complete explanatory history. The problem is that neither version of (C) will get all cases right.

Suppose that Nixon's missile system is indeterministic. When the button is pushed, there is some chance of a nuclear explosion and some chance that the signal

will fizzle out. In this case, we want to say that there might have been a nuclear explosion or the signal might have fizzled out if Nixon had pressed the button. Explosion-worlds and fizzling-worlds are equally close. But note that there are countless matters of particular post-antecedent fact that obtain both in the chancy-fizzling world and in our world, but not in the nuclear-explosion world (e.g. all the business as usual in Moscow, London or L.A.). Moreover, it would seem that these matters of fact are produced by the same factors in the chancy-fizzling world and in our world. And yet, since the world with chancy fizzling is no closer than the world with nuclear war, the extra post-antecedent similarities in the chancy-fizzling world must be irrelevant to closeness. The conclusion to draw is that, in order for the sharing of a fact between two worlds to contribute to the closeness between them, it is not sufficient that the *producing* explainers of this fact be the same in the two worlds. If we interpret the notion of sameness of explanation in the weaker way, as sameness of productive explanatory history, then the right-to-left ('if'-) direction of (C) will be false.

On the other hand, if we give the notion of sameness of explanation the stronger reading as sameness of complete explanatory history, then the left-to-right ('only if'-) direction will be false. This was brought to my attention by Jonathan Schaffer by means of an example of which I will here present a variant. Suppose that the king tosses a fair coin and it comes up heads. On the eve of the coin toss, a would-be assassin who is about to plant a bomb at the royal palace is poisoned by his enemy *x*. *x*'s action is an element of the complete explanatory history of the king's coin toss, since it prevented the assassin from killing the king and thereby interfering with the causal process that led up to the coin toss. Hence, in a world in which a different person, *y*, poisons the assassin, the complete explanatory history of the coin toss is different. And yet we want to say that the coin toss would have yielded the same outcome if the assassin had been poisoned by *y* rather than by *x*.

These considerations suggest that (C) can only be an approximation to the correct account of the phenomena considered in sections 3 and 4. More work on the details is needed. Attempts to develop a more refined version of the principle have convinced me that this is no easy task. It requires careful attention to large number of examples,

some of them quite complex. (I also happen to think that a correct account requires us to modify the similarity account as formulated in (0).) I am not sure how to tackle this problem. In this paper, I have merely tried to make it plausible that some principle in the neighbourhood of (C) is correct, and that this principle has a lot of explanatory power. I think that the foregoing considerations make this claim plausible even in the absence of an account that gets all the details right.

9. Conclusion

If the foregoing discussion is on the right track, then the concept of explanation ought to take centre stage in the theory of the truth-conditions of standard counterfactuals: Their truth-values depend crucially on the explanatory interrelations between different facts of our world. (C) is intended as a rough and approximate characterization of this dependence. As we saw above, it stands in need of refinement if it is to accommodate all relevant data.

I argued that a principle like (C) can explain a multitude of data, including the asymmetry of counterfactual dependence. I argue elsewhere that its explanatory power extends beyond that:

(i) In my (forthcoming) I argue that it might also be possible to use (C) to explain the special role of the laws of nature in the truth-conditions of counterfactuals, in particular the fact that laws have the power to support counterfactuals while accidental regularities do not.

(ii) Throughout this paper, I have used the possible-worlds theory (0) as my working account of the truth-conditions of counterfactuals, and (0) was perfectly adequate for this purpose. However, the theory faces a well-known problem that will ultimately require attention: If the antecedent of a counterfactual is impossible, then there are no possible antecedent-worlds, so that (0) makes the counterfactual vacuously true. But it seems implausible to me (and to many others) that all counterfactuals with impossible antecedents are true. (Consider, ‘If I were a hippopotamus, I would be a reptile.’ I believe that the antecedent of this counterfactual is impossible, but the counterfactual

does not seem true to me.) I am very sympathetic to a solution to this problem which Daniel Nolan, among others, has begun to develop: Simply allow impossible worlds to figure in your account of counterfactuals along with possible worlds. Impossible worlds are ordered with respect to their closeness to our world, just as possible worlds are. A counterfactual with impossible antecedent is true just in case its consequent is true in all the closest impossible worlds in which the antecedent is true. (The idea that there are impossible worlds is problematic on a Lewisian realist conception of worlds, but unproblematic if we think of worlds as abstract entities, such as sets of propositions.)

I believe that it should be an important principle of such an account that possible worlds are closer to our world than impossible worlds.[28] And I tentatively suggest in my (forthcoming) that it might be possible to use (C) to explain why the possible worlds are closer.

Appendix: *An alternative solution to the problem considered in section 5*

Consider again the example of the master law and the master parameter that I considered in section 5. In section 5, I argued that, if we accept that laws cannot have exceptions, then we either need to deny that (15) and (16) are true in the example, or we need to accept the consequence that (17) is true in the example. I recommended that we resolve this problem by allowing laws to have exceptions. Needless to say, I am aware that some readers might be unwilling to accept this solution. This appendix is addressed to such readers. I will outline an alternative way of resolving the problem that is consistent with the assumption that laws must be true without exception. This alternative strategy rests on a modification of both (0) and (C).

Assume, then, that there can be no exceptions to a law of nature. And suppose once again that there is a master parameter and a master law with the properties described in section 5, and consider a counterfactual that starts 'If I had scratched my nose a minute ago … .' (I will assume that the antecedent is false.) The closest

[28] Nolan (1997, p. 550) discusses this principle, which he calls 'Strangeness of Impossibility'.

antecedent-worlds must either be unlike the actual world throughout the pre-antecedent time, or they need to feature some small miracle. The strategy to be described will assume that they are like the actual world until shortly before the antecedent-time but contain small miracles. Given the present assumption that laws cannot have exceptions, a nose-scratching world with miracles must be one in which the master law is not a law. Accordingly, it must be a world in which the explanations of all matters of particular fact are different from what they are in the actual world. If we were to accept (C), we would need to conclude that no similarities in matters of particular fact between the closest nose-scratching worlds and the actual world can contribute to the closeness of these nose-scratching world, and that the closest nose-scratching worlds must therefore include worlds whose matters of particular fact are arbitrarily different from those of our world. Since this would lead us to back to (17), we need to revise (C) somehow. More specifically, we need to formulate a restricted version of (C) that permits us to say that similarities in matters of particular fact of the closest nose-scratching worlds contribute to closeness, despite the fact that the relevant matters of fact do not have quite the same explanations.

A suitable revision of (C) must preserve (C)'s capacity to explain the kind of intuition that motivated (C) in the first place, like those discussed in sections 3 and 4. If we want to find a suitable way of weakening (C), we need to find some principled way of distinguishing between examples like (9) ('If (Law of Gravitation) had not been a law, then events would still have at least approximately conformed to it') for which (C) yields the right predictions, and cases like the one considered in the last paragraph for which it does not. I think that one very natural thing to say is the following: One of the facts that, in the actual world, contribute to explaining why events conform to (Law of Gravitation) (namely the fact that (Law of Gravitation) is a law) is flatly inconsistent with the antecedent of (9). That is why, in the case of (9), the conformity of an antecedent-world to (Law of Gravitation) contributes nothing to its closeness, and (9) comes out false. Compare this with the example of the last paragraph. Although (given the assumption that miracles occur in the closest antecedent-worlds) the master law is not a law in the closest nose-scratching worlds, the fact that it is a law is not inconsistent with the antecedent ('I scratched my nose a

minute ago'). This is why similarities between antecedent-worlds and the actual world in matters which are actually explained by the lawhood of the master law can still matter to closeness.

This suggests that (C) is an over-generalization of (c). According to (c), if some matter of particular fact *f* obtains both in the actual world and in the antecedent-world w_A, this similarity is irrelevant to the closeness between the two worlds if different factors figure in the causal history of *f* in the actual world and in w_A. As a special case of this, the similarity is irrelevant if some of the factors that figure in the causal history of *f* in the actual world do not exist at all in w_A. And as a special case of *that*, the similarity is irrelevant if some of the factors that figure in the causal history of *f* in the actual world are inconsistent with the antecedent, and so cannot exist in *any* possible antecedent-world. The considerations of the preceding paragraph suggest that it is only this last, restricted principle that generalizes to explanation in general (causal and non-causal).

Before I can try to provide a rigorous formulation of this idea, I need to reconsider a presupposition that I have been making so far. I have assumed that there is *one* closeness relation that enters into the standard truth-conditions of all counterfactuals. That is controversial. Some philosophers have thought that different closeness relations enter into the standard truth-conditions of different counterfactuals if they have different antecedents.[29] I will now drop the controversial presupposition and allow the standard closeness ordering between worlds to be antecedent-relative. When speaking of the closeness relation that enters into the standard truth-conditions of a counterfactual with antecedent *A*, I will call it '*A*-closeness'. I propose that we replace (C) by something roughly like the following principle, which enshrines the idea adumbrated in the preceding two paragraphs:

(C*) Consider any counterfactual 'If *A* had been true, then *C* would have been true'. Let w_A be some antecedent-world. Suppose that a certain fact *Y* obtains both in the actual world and in w_A. This similarity

[29] Bennett, for example, mentions in (2003), pp. 298–301, that he used to hold this view. Also see Mårtensson (1999), Scts. 1.5.6, 2.8.4, 3.5.

between the worlds contributes to the *A*-closeness of w_A to the actual world if and only if

(a) the same factors figure in the causal history of *Y* in the actual world and in w_A; and

(b) in the actual world only facts that are consistent with the antecedent *A* contribute to explaining *Y*.

The 'only-if' (i.e. left-to-right) direction of (C*) imposes a constraint that a similarity between worlds needs to meet in order to contribute to closeness. Unlike the constraint imposed by (c), that imposed by (C*) is strong enough to explain our reluctance to accept (9) and (10). At the same time, it is weak enough to avoid the problem that the nose-scratching example presents for (C): Consider a well-behaved nose-scratching world *w*, that is, a nose-scratching world that diverges from ours by a small violation of the master law shortly before nose-scratching time and afterwards unfolds in accordance with the master law. Given the temporal asymmetry of causation, the only factors that figure in the actual causal histories of matters of particular pre-divergence fact are themselves located at times before the divergence, and thus also obtain in *w*. Hence, the matters of particular fact during the pre-divergence period have the same causal histories in the actual world and in *w*. Moreover, the only factors that contribute to explaining matters of particular pre-divergence fact in the actual world are the fact that the master law is a law and other matters of particular pre-divergence fact, and the obtaining of each of these explanatory factors is consistent with the antecedent. The pre-divergence similarities between the actual world and *w* thus satisfy conditions (a) and (b) of (C*). Hence, according to (C*), they contribute to closeness. In contrast to (C), (C*) is thus consistent with the assumption that the well-behaved antecedent-worlds are closer to the actual world than any other antecedent-worlds, and that things would not have been very different if I had scratched my nose.

An account that incorporates (C*) does not presuppose that the closeness ordering among worlds is antecedent-invariant, and it is in fact easy to see that according to (C*) it is not. Consider counterfactuals of the form,

(26) If the master law were not a law, then it would be the case that *q*.

By hypothesis, the lawhood of the master law contributes to explaining every matter of particular fact and the lawhood of every law (except the master law itself). But the fact that the master law is a law is inconsistent with the antecedent of (26). We can therefore conclude from (C*) that there are no similarities between an antecedent-world and the actual world in matters of particular fact or law that matter to closeness. This suggests that, as far as concerns the closeness relation that is relevant to the truth-conditions of (26), pretty much all worlds in which the master law fails to be a law are equally close. But now consider again

(27) If I had scratched my nose a minute ago, things would be very similar now.

On the version of my account that we are exploring in this appendix, the closest antecedent-worlds are the well-behaved ones. The well-behaved antecedent-worlds (which contain violations of the master law and in which the master law therefore fails to be a law) are thus closer than any other antecedent-worlds in which the master law fails to be a law. As concerns the closeness relation that enters into the truth-conditions of (27), not all worlds in which the master law fails to be a law are equally close. This means that the closeness relation relevant to (27) must be different from that which is relevant to (26).

As the considerations of this appendix show, for those who think that laws cannot have exceptions, (C*) must seem to have an important advantage over (C), given that it allows us to solve the problem considered in section 5. Needless to say, this advantage comes at a cost. Like Lewis, the proponent of (C*) needs to endorse the implausible claim that, if determinism is true, then the laws would have been different if I had scratched my nose. She also needs to reject as invalid some principles of inference that might intuitively appear to be valid, like the following rule of restricted hypothetical syllogism:

If it had been the case that *p*, then it would have been the case that *q*.
If it had been the case that *p* and *q*, then it might have been the case that *r*.
∴ If it had been the case that *p*, then it might have been the case that *r*.

This weakening of the logic of counterfactuals results from the antecedent-relativity of the closeness relation. We would ordinarily show that the foregoing inference principle is valid by roughly the following three-step argument:

1. Assume that the two premises are true. It follows that all the closest *p*-worlds are *q*-worlds and that some of the closest ($p \wedge q$)-worlds are *r*-worlds.
2. Since all the closest *p*-worlds are *q*-worlds, the closest ($p \wedge q$)-worlds must be identical with the closest *p*-worlds.
3. Hence, the fact that some of the closest ($p \wedge q$)-worlds are *r*-worlds entails that some of the closest *p*-worlds are *r*-worlds, so that the conclusion is true.

This argument rests on the implicit assumption that *p*-closeness is the same relation as ($p \wedge q$)-closeness, and this cannot be taken for granted if different closeness-orderings are relevant to counterfactuals with different antecedents. Similar problems arise for other inferences from counterfactual premises to a counterfactual conclusion if the premises and the conclusion have different antecedents. The assumption that the premises are true imposes constraints on the closeness ordering of worlds, and an attempt at validation will usually proceed by showing that the only closeness orderings that meet all of these constraints also make the conclusion true. But this strategy obviously works only if the closeness ordering relevant to the conclusion is the same as the closeness orderings relevant to the individual premises. And we have no right to assume that this is so if the premises and the conclusion have different antecedents and the closeness ordering is antecedent-relative.

If we accept (C*) and endorse the view of counterfactuals outlined in this appendix, then counterexamples to the above inference rule of restricted hypothetical syllogism are readily forthcoming. Suppose that there is a master law with the features already described. On the view expounded in this appendix, the master law would not have been a law if I had scratched my nose. Moreover, if the master law had not been a law and I had scratched my nose, then pretty much anything might have been case. (Most facts in the actual world are explained, in part, by the lawhood of the master law, and the lawhood of this law is inconsistent with the antecedent.) And yet it is not true that anything might have been the case if I had scratched my nose (most things would have been the same).

(C*), with its two different clauses for causation and explanation on its right-hand side, displays an unpleasant lack of unity. Suppose that some fact *f* obtains at *w* and also at the *A*-world w_A. In order for this similarity to count towards *A*-closeness between *w* and w_A, *all* the factors that figure in the causal history of *f* in *w* must also obtain at w_A. By contrast, it is not true that all those factors that contribute to explaining *f* in *w* without figuring in *f*'s causal history in *w* must also obtain at w_A in order for the similarity to contribute to *A*-closeness. Some of the latter factors may be absent, provided they are not inconsistent with the antecedent *A*. According to (C*), then, causation and non-causal explanation enter into the truth-conditions of counterfactuals in different ways. Would a unified account of the role of explanation in the truth-conditions not be more satisfactory?

I sympathize with this uneasiness. Perhaps the best way in which one can attempt to allay the worry is by trying to give a functional explanation of why our counterfactual reasoning is governed by a gerrymandered rule like (C*). One may try to explain why the notion of causation and the concept of non-causal explanation play different roles in the truth-conditions of counterfactuals by arguing that the roles of the two notions have different sources; more concretely: that they have their sources in different functions that counterfactuals serve. Let me explain. Counterfactual thoughts have a number of important functions in our thinking. We use them, for instance, when trying to explain things ('It is only because of you that he is not here. He would have come if you had not been so nasty to him.'), and when deciding what

to do ('If I were to call him now, he would be able to come tomorrow.'). For each of these two purposes, we can say: In order for counterfactuals to serve this purpose adequately, their truth-conditions must be governed by a closeness relation that satisfies such-and-such a condition. Each of the two purposes imposes a different constraint on the closeness relation. In order for the counterfactual connective to serve *both* purposes, it must conform to *both* constraints. And it could be that the two clauses in (C*) are reflections of these two different constraints. Perhaps the clause relating to causal explanation reflects a constraint that has its source in our use of counterfactuals in decision-making, whereas the clause relating to explanation in general reflects a constraint that arises from our use of counterfactuals in evaluating claims about explanation. It remains a task for another occasion to expand on this idea.[30]

References

Adams, Ernest 1975: *The Logic of Conditionals*. Dordrecht: Reidel.

Asquith, Peter and P. Kitcher (eds.) 1984: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1984*. East Lansing, MI: Philosophy of Science Association.

Bennett, Jonathan 1974: 'Review of David Lewis, *Counterfactuals*'. *Canadian Journal of Philosophy*, 4, pp. 381–402.

—1984: 'Counterfactuals and Temporal Direction'. *Philosophical Review*, 93, pp. 57–91.

—2003: *A Philosophical Guide to Conditionals*. Oxford: Clarendon.

Collins, John, N. Hall, and L. Paul (eds.) 2004: *Causation and Counterfactuals*. Cambridge, MA: Bradford Book / MIT Press.

[30] I am grateful to many philosophers for their comments on various bits and pieces of this material, including Gordon Belot, Karen Bennett, John Burgess, Jeremy Butterfield, Cian Dorr, Ant Eagle, Adam Elga, Delia Graff, Gilbert Harman, Thomas Hofweber, Mark Johnston, Marc Lange, Stephan Leuenberger, Ram Neta, Jim Pryor, Peter Railton, Geoffrey Sayre-McCord, Jonathan Schaffer, Kieran Setiya, Fritz Warfield, Brian Weatherson, an anonymous referee for *Mind*, as well to the audiences of talks I gave at Princeton, Cornell, Pittsburgh, Michigan and the University of North Carolina at Chapel Hill, and the members of a graduate seminar on the metaphysics of modality that I gave at the University of Michigan at Ann Arbor in the Fall Term of 2005. My greatest debt is to my dissertation advisor Gideon Rosen, who discussed numerous previous drafts of this paper and of related material and made many helpful comments and suggestions.

Dowe, Phil and P. Noordhof (eds.) 2003: *Causation and Counterfactuals*. London: Routledge.

Edgington, Dorothy 2003: 'Counterfactuals and the Benefit of Hindsight', in Dowe and Noordhof 2003, pp. 12–27.

Elga, Adam 2000: 'Statistical Mechanics and the Asymmetry of Counterfactual Dependence'. *Philosophy of Science*, suppl. vol. 68, pp. 313–24.

Fine, Kit 1975: 'Review of Lewis's *Counterfactuals*'. *Mind*, 84, pp. 451–8.

Hall, Ned 2004: 'Two Concepts of Causation', in Collins, Hall and Paul 2004, pp. 225–76.

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

Johnson, David 1991: 'Induction and Modality'. *Philosophical Review*, 100, pp. 399–430.

Kment, Boris. Forthcoming: 'The Closeness Account of Necessity'. Forthcoming in *Philosophical Perspectives*.

Lange, Marc 2000: *Natural Laws in Scientific Practice*. New York: Oxford University Press.

Lewis, David 1973a: *Counterfactuals*. Cambridge, MA: Harvard University Press.

—1973b: 'Causation', in Lewis 1986b, pp. 159–213. Originally published in *Journal of Philosophy*, 70, pp. 556–67.

—1979: 'Counterfactual Dependence and Time's Arrow', in Lewis 1986b, pp. 32–52. Originally published in *Noûs*, 13, pp. 455–76.

—1986a: 'Postscripts to "Counterfactual Dependence and Time's Arrow"', in Lewis 1986b, pp. 52–66.

—1986b: *Philosophical Papers*. New York / Oxford: Oxford University Press, vol. ii.

Mårtensson, Johan 1999: *Subjunctive conditionals and Time*. Göteborg: Kompendiet.

Nolan, Daniel 1997: 'Impossible Worlds: A modest approach'. *Notre Dame Journal of Formal Logic*, 38, pp. 535–73.

Nute, Donald 1980: *Topics in Conditional Logic*. Dordrecht: Reidel.

Salmon, Wesley 1984: 'Scientific Explanation: Three General Conceptions', in: Asquith and Kitcher (eds.), 1984, pp. 293–305.

Slote, Michael 1978: ‘Time in Counterfactuals’. *Philosophical Review*, 87, pp. 3–27.

Stalnaker, Robert 1968: ‘A Theory of Conditionals’. *Studies in Logical Theory*, *American Philosophical Quarterly* Monograph Series, 2. Oxford: Blackwell, pp. 98–112.

Tichý, Pavel 1976: ‘A counterexample to the Stalnaker-Lewis analysis of counterfactuals’. *Philosophical Studies*, 29, pp. 271–3.

Wasserman, Ryan. Forthcoming: ‘The Future Similarity Objection Revisited’. Forthcoming in *Synthese*.