

The Content of Counterfactuals and their Role in Explanation

or

The Benefit of Hindsight

Dorothy Edgington

1. I am driving to the airport to catch a 9 o'clock flight to Paris. The car breaks down on the motorway. I sit there, gnashing my teeth, waiting for the breakdown service. 9 o'clock passes: I've missed my flight. More time passes. "If I had caught the plane, I would have been half way to Paris by now", I say to the repairman who eventually shows up. "Which flight were you on?" he asks. I tell him. "Well you're wrong", he says. "I was listening to the radio. It crashed. If you had caught that plane, you would be dead by now".

With a bit more elaboration, which it will get in due course, this story is an example of a kind which creates a difficulty for all well-known theories of counterfactuals, and for a view I held. The problem is not new--it was mentioned in the 1970s--but it has received relatively little discussion until recently. Its force was impressed upon me by the work of Stephen Barker (1998, 1999). I had not entirely ignored it before, mentioning it in passing (Edgington 1995 pp. 257-8), in criticism of David Lewis (1979). Yet later in the same article (§§ 8 and 10, especially pp. 320-1), when saying what we aim at in counterfactual judgements--when such a judgement is objectively correct--I had forgotten about this sort of example, and so got things wrong. I shall try to rectify that. And I shall try to explain why we assess counterfactuals the way we do, in the context of an answer to the question: what do we need counterfactual judgements for? What purpose do they serve for us and why

does it matter to us to get them right? What else goes wrong for us if we get counterfactuals wrong?

Before discussing the difficulty, I shall sketch the "standard picture", common to various theories, and then my version of this standard picture. I shall only be concerned with counterfactuals whose antecedent and consequent are about particular states of affairs holding at particular times, and those in which the consequent-time is later than the antecedent-time. Of course a theory needs to be more general than this, but that will not concern me here.

2. *The standard picture: Goodman and Lewis* The problem of counterfactuals has always been: what are the rules of the game? (There is also the question: why play this game? And if we answer that, which rules are appropriate to our purposes?) You suppose that something *A* had been true, something that is, often you know, actually false. You wonder whether, given that supposition, something else *C* would have been true. In trying to settle the matter, you need to rely on some actual facts, and let other actual facts go by the board with the supposition that *A*. What determines what you can hang on to, and what you must give up? Goodman (1955) gave us the form of a theory: ' $A \Rightarrow C$ ' is true iff *C* is deducible from *A* together with the laws of nature together with facts which are cotenable with *A*. But he gave up when trying to specify which facts are cotenable with *A*. Something is not cotenable with *A* iff it would not be true of *A* were true. This is unacceptably circular.

At first, Lewis's theory looked very different from the Goodman-style theories it succeeded. Call an *A*-world a world in which *A* is true. $A \Rightarrow C$ is true iff *C* is true in all "closest" *A*-worlds, i.e., all *A*-worlds which overall most resemble the actual world. Many

readers of Lewis's book (1973) assumed that ordinary common-or-garden standards of similarity were being invoked. Reviewers of the book, and others, pointed out that this doesn't work (see Bennett 1974, Fine 1995). By ordinary standards of similarity, the questions 'What would have happened if it had been the case that *A*?' and 'What is true in all *A*-worlds most similar to the actual world?' can get different answers. Any counterfactual of the form 'If *A*, then things would have been very different from the way they actually are' presents a difficulty. Fine's example, discussed by Lewis: if Nixon had pressed the button in 1974, there would have been a nuclear holocaust. But in the world most like the actual world in which Nixon pressed the button, nothing untoward happened. Lewis labels this the 'future similarity objection'. The objection is not answered just by discounting similarity after the consequent-time: its effect can arise from relying on similarity between antecedent-time and consequent-time. If Hitler had died in infancy, things would have been very different in the 1930s and 1940s. In the worlds in which Hitler died in infancy which most resemble the actual world up to the 1940s, however, some other child grows up to play a virtually identical Hitler-like role. In replying to these objections Lewis (1979) was more explicit about which factors count towards closeness on the 'standard resolution' of the vagueness of the notion of similarity. His criteria are stated in more general terms to cover not only sequential counterfactuals about particular facts, but they have this consequence for the latter: the closest *A*-worlds are those with pasts identical to the actual world, up to shortly before the antecedent-time, when we need to deviate just enough to get the antecedent true. (Call the point of deviation the fork.) The closest *A*-worlds obey the laws of nature of the actual world, except insofar as we may need a small, inconspicuous deviation to get us to depart from the actual world at all. There is no

deviation from the laws of nature after the fork. And that is all, or almost all. After we have deviated from perfect match, at the time of the fork, "It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly" (1986 p. 48). (We shall return to this disjunction.) It is just the laws that we rely upon, after the fork.

This is what I shall call the standard picture. Note that the refinement of Lewis's theory brings it closer to Goodman's, with a strong hint about which facts are cotenable with the antecedent. Instead of saying that *C* is true in all *A*-worlds with the same particular facts up until the time of the fork and the same laws thereafter, we could say that *C* is deducible from *A* + laws + facts up to the time of the fork. Differences may show up if we consider a wider range of counterfactuals. Problems might arise about the short "transition period" from the fork to *A*; but they usually don't. To a first approximation, they deliver the same picture. Others (e.g. Michael Slote (1978)) have given Goodmanian theories somewhat along these lines.

3. *The standard picture: probabilistic version* The view of counterfactuals I held (inspired by Ernest Adams (1975)) can be seen as a small modification of the standard picture described above. The truth conditions of the Lewis-Goodman type are, in my view, too strong. They make it too easy for a counterfactual to be plain false. Very many believable counterfactuals--possibly all or almost all the contingent counterfactuals we ever utter--could turn out to be downright false, on this version. I give three reasons:

(1) Indeterminism. Suppose that all or many fundamental laws of nature are indeterministic: they may operate so as to make a certain outcome extremely probable given

certain conditions, but not certain. Then no or few ordinary propositions will be deducible from antecedent + laws + cotenable facts. Mutatis mutandis, no or few ordinary propositions will be true in *all* closest antecedent-worlds. To the extent that this is so, ordinary counterfactuals will be false, according to these truth conditions. If we believe that this is so, we should, on this theory, have no confidence at all in any counterfactual. I submit that instead, if we believe that this is so, we should be less than completely certain, but are entitled still to be pretty confident--very close to certain, that if you had lit the gas, the water would have boiled, and so forth.

(2) Determinism. Even if we do live in a deterministic world, we do not live in a crudely deterministic world. Our ordinary run-of-the-mill antecedents are not normally specific enough to be fed into deterministic laws. Even if coin-tossing is a deterministic process, no deterministic conclusion comes from the counterfactual supposition that you had tossed the coin, but only from a supposition of how *exactly down to the minutest detail* you tossed it, together with, down to the minutest detail, the positions of the air molecules, etc.. An example I have used to illustrate both these cases: a dog almost always, but not quite always, attacks and bites when strangers approach. We can detect no difference between the cases in which it does and those in which it doesn't. Assume either there's some indeterminism involved; or else, if there isn't, the outcome depends in some immensely subtle way on the manner of approach. I say "I didn't approach, because I'm pretty sure that the dog would have bitten me if I had approached". But on the Goodman's theory, it is certainly false that if I had approached, I would have been bitten--either because of indeterminism, or, because the mere (coarse) supposition that I approached, together with all cotenable facts and laws, does not entail that I was bitten. And what is

certainly false is not something of which you should be close to certain.

The result is the same on Lewis's theory: assuming either indeterminism or fine-grained determinism, in almost all, but not quite all, close worlds in which I approach, I am bitten. That leaves the counterfactual clearly false.

(3) The vocabulary in which the antecedent and consequent are couched may not be suitable for subsumption under the deterministic laws. This is particularly relevant to the countless counterfactuals we accept and assert about our own and others' mental lives. "If I had received your invitation yesterday, I would have accepted". Take the Davidsonian view. Even if determinism is true, these are not the categories which belong with the deterministic laws. Again, the assumption of the antecedent, together with other facts and laws, does not enable you to deduce the consequent. All such counterfactuals are false, on Goodman's and Lewis's theory. Whereas, it seems to me, our confidence (perhaps short of certainty) in counterfactuals like these: "If you'd invited me yesterday, I would have accepted", "If Mary had asked John to do the shopping, he would have done so", "If Bill had been in London, he would have been in touch", does not depend upon our accepting that there are deterministic laws connecting consequent to antecedent and other relevant facts. Here is a perfectly ordinary use of a counterfactual: "They're not at home; for the lights are off; and *if they had been at home, the lights would have been on*". (The example is used by Adams.) You might be close to certain of the conditional, even if you are sure that their sitting in the dark is *not* inconsistent with the laws and cotenable facts. To repeat: on Goodman's theory, if you are sure that the consequent isn't entailed by laws etc, you should be sure that the counterfactual is false.

I am *not* recommending that we say instead that a counterfactual is true iff the

consequent is very probable given the antecedent, laws and cotenable facts. That won't work. Suppose we did say that. Suppose I know that it is indeed very probable that *C* would have been true if *A* had been true--say 95% probable; so I should be certain that it is true. So I should be certain that if *A*, *C*. But I'm not. I'm only close to certain. Suppose I know that it is not very probable that if *A*, *C*--it's around 50-50. Then I should be certain that it is not true. So I should have zero confidence that if *A*, *C*. But I don't: I think it is about 50-50 that *C* would have happened if *A* had. I'm suggesting instead that we simply stick with the appropriate conditional probability--the conditional probability of *C* given *A* *at the time of the fork*, as a measure of the acceptability of the counterfactual. You ask: how likely *was* it, *then*, that *C* would have happened if *A* had? One way of looking at it: consider all the Lewisian closest *A*-worlds. Suppose for simplicity that you have divided them into a finite number of equi-probable clumps in a suitable way. Then the question is, in what proportion of the clumps is *C* true? Whereas for Lewis, unless *C* is true at all the clumps, the counterfactual is plain false.

This view also fits with my view of indicative conditionals, and in particular, vindicates a nice relation between forward-looking "will"-conditionals and counterfactual "would"-conditionals. We believe an indicative conditional to the extent that we think the consequent is probable on the supposition of the antecedent. For many forward-looking "will"-conditionals, there is an objectively correct opinion to have: the objective chance of *C* given *A*. A boring and easy example: you are to pick a ball at random from a bag in which 90% of the red balls have black spots. What should you think about the conditional 'If I pick a red ball it will have a black spot'? You should be 90% confident that if you pick a red ball, it will have a black spot. That is the right, unimprovable opinion, at least

before you pick. Suppose you do pick a red ball. Then this conditional probability will change--collapse--to 1 or 0. Suppose you don't pick a red ball. Then it doesn't collapse. It's 90% likely that if you *had* picked a red ball, it would have had a black spot. And there it remains, unalterable forevermore (or so I thought). Even God can't better that judgement.

In central cases, "would"-conditionals and "will"-conditionals differ merely in a temporal way: the same conditional thought can be expressed now with a "will", later with a "would". I say "Don't go in there; if you go in you will be hurt." You look sceptical but stay outside, and there is a loud bang as the ceiling collapses. "You see", I say, "I was right: if you had gone in, you would have been hurt. *I told you so*". Or, if there is no loud bang and the ceiling doesn't collapse, "I was wrong; I thought the ceiling was about to collapse; I thought you would have been hurt if you had gone in". "If they're here by eight, we'll eat at nine" is rephrased hungrily at ten "If they had been here by eight, we would have eaten at nine". I change my travel plans on being told "If you travel on Friday, it will cost you £20 extra". I discover I was misinformed: if I had travelled on Friday, it would not have cost me extra. And so on. Your present "would haves" agree with your present opinion about the acceptability of the corresponding earlier "will".

4. *The Problem* Return, at last, to the plane crash. Stipulate that a chance event, not predictable in advance, brought down the plane. Everyone aboard was killed. Indeed, there was no chance that anyone on board would survive. At the time of take-off, this plane was not relevantly different, with respect to safety, from any other normal plane: there was an extremely small but non-zero chance that some such accident would occur--due to freak weather conditions, or freak electrical or mechanical faults (or combinations thereof), or a

freak heart attack or attacks on the part of those in control.

Is the repairman's remark correct? Well, perhaps not if, for example, some subtle feature of the distribution of weight in the plane played some causal role in the antecedents of the crash--a feature which might well have been different, had I been on board. But if, as is more likely, my absence from the plane had no effect on the aetiology of the crash, it is surely correct.

The first mention of an example like this in print is in a footnote at the end of a paper by Slote (1978), and is attributed to Sydney Morgenbesser. It is simply "If I had bet on heads, I would have won", said of a presumed indeterministic coin-toss which landed heads. Similarly, any week after a lottery draw, I'm right, it seems, to say "If I had chosen numbers 45 67 I would have won".

Slote says in this footnote: "I know of no theory of counterfactuals which can adequately explain why such a statement seems natural and correct. But perhaps it simply *isn't* correct, and the correct retort is `no, you're wrong; if I had bet (heads), the coin might have come up differently, and (so) I might have lost--assuming the coin was random'" (p. 27).

This, I think, is wishful thinking³ (wishful philosophical thinking, that is: the example refutes the thesis of Slote's paper). Consider: you are watching a lottery draw on television and to your dismay your arch business rival wins a prize--not a big enough prize for him to abandon his business, but big enough for him to put you out of yours. If Slote's

³At least if the story is told in an appropriate way. As with the plane crash, the betting story is sensitive to whether my saying "Heads" might have influenced the manner in which the coin was tossed. To avoid this possibility, let the tossing happen in one room, and I write "Heads", "Tails" or "No bet" on a piece of paper in another room.

suggested "retort" were correct, so would this be: you say to yourself, "If I had scratched my nose a minute ago, he very probably would have lost. What a pity I didn't scratch my nose!"⁴

We can do better than just to appeal to intuitions. In fact, the appeal to intuitions is compelling, I think. But it leaves hanging the question of *why* our counterfactual thought-experiments are conducted in this manner. In the final section of the paper, I try to show how the intuitive response to these examples is the one that fits the use we make of these judgements.

(Note: you have to at least countenance the possibility of indeterminism for these examples to be a problem for the standard view. It seems to me (and to Lewis) that a decent theory of counterfactuals should cater for that possibility. But for someone who thinks, on something like *a priori* grounds, that determinism must be true, the standard view is not in trouble: sufficient causes of the plane crash were there back before the fork.)

A few more remarks about the plane crash. First, it was not essential to the story of that the crash was such that those on board had a 100% chance of being killed. Perhaps there were a few survivors. Perhaps there was about a 90% chance of being killed if on board. Even so, I will think, "it's very likely that I would have been killed if I had been on board"--unless I can tell a special story about my abnormal powers of survival.

Second, as mentioned above, there can be mixed cases where there is some chance that my presence on the plane would have altered conditions in a way to prevent the crash, and some chance that it would not have interfered with the crash. For a purer example, consider a coin toss. Case 1: I decline to bet. It lands heads. Assume no causal interference.

⁴Here I borrow from David Johnson (1991), one of the few discussions of this problem.

If I had bet on heads I would have won.

Case 2: there is a cheat around. He has a little gizmo in the palm of his hand. When someone bets heads, he presses it, it sends out a magnetic pulse or whatever, which prevents the coin landing heads. In this case, if I had bet on heads, the coin would not have landed heads.

Case 3: we have a more sophisticated cheat. After all, suspicion would be aroused if coins *never* landed heads when people bet on heads. He does not trust himself to randomise. The device does it for him. He always presses it when someone bets heads. There's a 90% chance that it does nothing, and a 10% chance that it prevents the coin from landing heads. I didn't bet. The coin lands heads. If I had bet on heads, it's 90% likely that I would have won; for there was a 90% chance of no causal interference, and a 10% chance that the coin would have been prevented from landing heads. Note that here too, the way the coin *actually landed* carries weight in assessing the counterfactual. (Examples like this are discussed in Barker 1999.)

Finally, let me stress again the crucial role of *causal* independence. As a fantasy, imagine that the crash has this genesis: the devil spins a spinner, which has a one-in-a-million chance of landing in the space designated 'crash'. It does land there. There is a crash. If I had caught the plane I would be dead. Now suppose that the devil has two identical spinners, and some rule for deciding which to use which has the consequence that he will spin one if I'm on the plane, the other if I'm not. Although the chances are initially just the same, in this case, if I had caught the plane, very probably it would not have crashed!

The problem for the standard view, in either version, is, of course, that actual facts

after the time of the fork can be crucially important to the assessment of counterfactuals.

5. *The Problem for Lewis* "It is of little or no importance to secure approximate similarity of particular fact, even in matters which concern us greatly" says Lewis (1986, p. 48). That is, after the fork when we no longer have perfect match, similarity of particular fact is of little or no importance. The nearest he gets to addressing our problem is in the parenthetical remark which follows: "(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. Tichy (1976) and Jackson (1977) give cases which appear to come out right ... only if approximate similarities count for nothing; but Morgenbesser ... has given a case which appears to go the other way ...)". That is all he says and he has never, as far as I know, returned to the problem. The Tichy example is this: when Fred goes out and it's raining, he always takes his hat. When he goes out and it's not raining, it's a random 50-50 whether he takes his hat. On this occasion, it's raining and he takes his hat. Consider 'If it had not been raining, he would have taken his hat'. The fine-weather world in which he does take his hat resembles the actual world more than the fine-weather world in which he does not take his hat resembles the actual world. But this, Lewis rightly wants to say, counts for nothing: the counterfactual is not clearly true. (For Lewis, it's clearly false; for me, it's 50-50.) This suggests the no-weight picture has to be the right one.

Many examples go the other way: if I'd bet on heads, I would have won; if I had bought these shares a year ago, I would be rich; if I had left 5 minutes earlier, I would have avoided the accident; if I had got up 5 minutes earlier, the result of the Australian General Election would have been just the same.

I pick a coin from a bowl of coins, toss it, and it lands heads. It would be wrong to claim that if I had picked a different coin, it too would have landed heads. But it would be absurd to deny that if Frank in Australia had scratched his nose a moment or two earlier, the coin I picked, and tossed, which actually landed heads, would still have landed heads. The difficulty for Lewis is distinguishing these cases.

Another such pair. Guerrilla warfare in an imaginary country. The guerrilla leader is hiding in a certain village. Government troops have a range of missiles aimed at the village. These devices are indeterministic, and each has a chance of, say, 90% of firing when activated. News having arrived of the need to deploy troops elsewhere, only one missile is to be set off. The General chooses a missile, which is activated. It fizzles out. No harm is done.

"We were lucky", says a potential victim later. "Had the General chosen a different missile, we might well be dead." His companion, versed in Lewis's early work on counterfactuals, taking "similarity" in an intuitive way, demurs. "We were lucky that *it* didn't go off", he says, "but your relief is misplaced. In the world most like the actual world in which he chose a different missile, it fizzled out too, right? That is to say, if he had chosen a different missile, it too would have fizzled out." The silliness of this suggests that Lewis should say "approximate similarity counts for nothing".

Version 2 of the story: two inhabitants of the village are delayed on their way home because they notice a sheep caught in a cactus, and it takes them a while to free it. The scenario is as before, but let me lower the chance each missile has of firing, to about 25%. This time, the missile does fire. They hear it in the distance. When they get back, they meet havoc and destruction. "If we hadn't noticed the sheep, we'd probably be dead now",

says one. His companion, versed in Lewis's later work and the reasons for saying "approximate similarity counts for nothing" demurs: "Consider the possible world which deviated from the actual one at the time we noticed the sheep: the missile (or its counterpart), in that world, had only a 25% chance of going off. So if we hadn't noticed the sheep, it's 75% likely that the disaster would not have occurred. What a pity we noticed the sheep! If we hadn't, probably, all would be well".

I don't see how Lewis can handle these examples without appealing to the notion of causal independence. Whether Fred wears his hat is not causally independent of the weather. Picking another coin or missile begins a different casual process. But the outcome for this coin or missile that was picked, is causally independent of someone's scratching his nose in Australia, or the antics of the sheep. As Lewis wants to explain causal dependence and independence in terms of counterfactuals, this is a problem for him.

It might be thought that the standard picture gives the conditions for causation, even if it doesn't always agree with our counterfactual judgements. But the problem cases seem to show that the standard picture gives the wrong conditions for causation--at least when we allow causation to be indeterministic, as Lewis does, and as do all who pursue this approach to causation. Lewis's account went like this: c directly caused e iff c occurred, e occurred and the actual chance, immediately after c occurs, that e will occur, is significantly higher than the chance of e occurring in the absence of c . Suppose I'm facing a machine which emits particles. I snap my fingers. Immediately afterwards, the chance that a particle is emitted reaches 100%. Are these events causally related? We have to assess the counterfactual "If I had not snapped my fingers, the chance would have been much less than 100% that a particle be emitted". To assess this, according to the standard picture, we go

back to the time of the fork, shortly before I snapped my fingers; we appeal to the actual laws of nature but not to particular facts thereafter; and we ask what is the chance, at a time just after the antecedent time, that a particle be emitted. It might be very low. So the standard picture delivers the wrong answer. (The example is Barker's (unpublished).) Of course we want to say: the particle would have been emitted even if I had not snapped my fingers. But that *rests on* a judgement of casual independence.

In addition to the point about the need to appeal to causation, I suggest that Lewis's modal realism makes it hard for him to put weight on the distinction that matters here. A die is tossed, and lands six. We can't infer that if another die had been tossed instead, it would have landed six. But we can infer that if Frank on the other side of the world had scratched his nose, this die would still have landed six. But for Lewis, this latter question is a question about whether, in all sufficiently close possible worlds in which counterpart Freds scratch their noses, counterpart dice in those worlds land six. It is hard to see what would *ground* the right answer, when it's put Lewis's way.

6. *Handling the Problem* If we give up on the idea of explaining causation in terms of counterfactuals (or never had that idea in the first place), it is not too hard to see how the standard picture needs to be amended to handle these examples. When we assess a counterfactual, we may need to take into account the way the world actually rolls on, after the fork, in ways which are causally independent of our antecedent. A Lewis-style account would go thus: consider those *A*-worlds which (a) depart from the actual world shortly before the time of $\neg A$, at an inconspicuous fork; (b) thereafter obey the actual laws of nature; and (c) share with the actual world subsequent particular facts which are causally

independent of $\neg A$, up to the time of the consequent. A counterfactual $A \Rightarrow C$ is true iff C is true at all such worlds. A Goodman-style story will say: the cotenable facts are (a) those up to shortly before the time of $\neg A$; (b) the laws of nature; (c) any subsequent fact, up to the time of the consequent, which is causally independent of $\neg A$ (i.e. whose causal history does not go through $\neg A$). $A \Rightarrow C$ is true iff C is entailed by A and the cotenable facts. I would subject both to probabilistic amendment, as before. Instead of Lewis's truth condition I would say (very crudely and roughly) that $A \Rightarrow C$ is probable to the extent that C is true in most of those A -worlds. Instead of Goodman's, $A \Rightarrow C$ is probable to the extent that the chance is high, at the time of the fork, of C given A , the laws and the cotenable facts. The objectively correct value to assign to such a counterfactual is not (or not always) the conditional chance of C given A at the time of the fork; but the conditional chance, at that time, of C given $A \& S$ where S is a conjunction of those facts concerning the time between antecedent and consequent which are (a) causally independent of the antecedent, and (b) affect the chance of the consequent.

At first sight this is a rather strange probabilistic animal, but it is a *bona fide* conditional probability. Think of it this way. Go to a time just before $\neg A$, and consider the chance then of C given A . There is a future of branching paths, some of which lead to C and some of which lead to $\neg C$, and with other possible intervening events along the various paths. For any actual fact S causally independent of A which affects the chance of C , cross out the $\neg S$ paths; and recalculate the chance of C given A on that basis.

This is what we aim at. Of course, we often don't know enough to get it right. Interesting questions arise about how best to estimate such a thing, in states of imperfect information. I won't go into that here.

What happens to the pleasing view of the relation between forward-looking "will"-conditionals and retrospective "would"-conditionals, on this amended view? With hindsight, I think that if I had caught the plane, I would have been killed; that if I had bet on heads, I would have won. But there was no reason to think beforehand that if I catch the plane, I will be killed, or if I bet on heads, I will win.

Consider this, however: about the plane crash, a friend has a powerful hunch, *or* has some erroneous reasons, *or* has some good-but-Gettier-like reasons, for thinking this plane will crash. "Don't take it!" he says. "If you catch that plane, you'll be killed". I shrug this off as irrational advice, which it is. I miss the plane. It crashes. "My God, he was right!", I say, on hearing the news. "If I'd caught the plane, I would have been killed!". Similarly, if someone tells me that if I choose ticket number 65 87 92 ... I will win; or that if I bet on heads, I will win; or that if I buy these shares, I will become rich; and so on. Even if not rationally grounded, these unfulfilled conditionals are vindicated. The case for the temporal relation between wills and woulds remains: one is right iff the other is. The hindsightful counterfactual vindicates the earlier "will", even if the "will" was not justified at the time.

We are familiar with the thought that rationally held beliefs may turn out false, and conversely, something which there is no reason to believe may turn out true. "Right belief" admits of two readings: rational belief, and true belief. If that were my story, there would be no novelty or mystery. But that is not my story. Counterfactuals, like other conditionals, are believed to the extent that a certain conditional probability is judged to be high, and that is not the probability of the truth of a proposition. The right value to assign to them is given by a certain conditional probability, not a truth value. It may be 1 or 0, but it need not be.

The fraudulent fortune-teller, gazing into her crystal ball, says "It's not altogether clear, but I'm pretty sure that if you fly this week, you will be killed". I miss my plane. It crashes. About 90% of those on board are killed. "My God she was right!", I say, "It was very likely that I would have been killed, had I caught that plane". Lucky guesses are sometimes right, and this was one. The value to be assigned to the hindsightful counterfactual trumps the most rational value to be assigned to the forward-looking indicative.

7. *What are counterfactuals for?* The question is pressing. Why do we evaluate counterfactuals the way we do? What would go wrong for us if we chose to evaluate them in some other way, e.g. according to the "standard picture"? The question deserves more attention than it has had in the vast literature on counterfactuals. I don't pretend to an exhaustive answer, but highlight some important aspects of their use.⁵

We use counterfactuals in empirical inferences to conclusions about what is actually the case. We need to try to get them right, in order to avoid, as much as possible, arriving at wrong conclusions about what is the case. I shall concentrate on two such forms of inference. There may be more, but these, I think, are central. Some examples:

(1a) You are driving, of an evening, in the dark, close to the house of some friends, and have considered paying a visit. You turn the corner. "They're not at home", you say. "For the lights are off. And if they had been at home the lights would have been on."

(1b) "It's not a problem with the liver", says the doctor". For the blood test was

⁵These thoughts owe a great deal to Ernest Adams: chapter 4 of *The logic of Conditionals*, and his article 'On the Rightness of Certain Counterfactuals', *Pacific Philosophical Quarterly* 1993. No one else, to my knowledge, has investigated this aspect of the use of counterfactuals.

normal. And if it had been a liver problem, it would have been [such-and-such]". Call these types of inference "counterfactual modus tollens".

(2a) A patient is brought to hospital in a coma. "I think he must have taken arsenic", says the doctor, after examination. "For he has [such-and-such] symptoms. And these are just the symptoms he would have if he had taken arsenic." (Note that in calling the conditional in this inference "counterfactual" we are using the label as a proper name for a form of conditional. There is nothing literally counterfactual about it. The example comes from Anderson (1951).)

(2b) The prison warden on his rounds says "I think a prisoner escaped from that window. For the flowers below are all squashed. And they would have been squashed if he had jumped from there." Call this style of inference "inference to the best explanation".

So we have two forms:

- (1) H. Because, E; and if it had not been the case that H, it would not have been the case that E;
- (2) H. Because, E; and if it had been the case that H, it would have been the case that E.

Neither of these forms of inference is valid. They are defeasible forms of empirical reasoning. This is obvious in the case of (2). (2a) could be defeated by pointing out that although these are indeed the symptoms he would have if he had taken arsenic, they are also the symptoms he would have if he had not taken arsenic but was, say, epileptic. (2b)

could be defeated by pointing out that the flowers would also have been damaged if a prisoner had not escaped but there had been a game of football, or a dog fight.

The same is true of (1). (1) is closer to valid in the following sense. If each premiss is certain, the conclusion is certain. But contingent conditional premisses of this kind are rarely certain, and we need to use them when they are less than certain. And an argument deserving the appellation 'valid' is such that if both premisses are close to certain, so is the conclusion (not quite so close, perhaps, but still close). That is a property demonstrably had by all paradigmatically valid arguments. (1) does not have this property. It can be defeated thus: "I agree that it was indeed very likely that we would find the lights on, if they were at home; but it was also very likely that we would find the lights on, if they were not at home; for they have the deeply engrained practice of leaving the lights on when they go out at night. So there must be some other explanation of the lights' being out. Perhaps there's a power cut; or they have gone to bed early".

When we see what defeats them, we see that the two forms of inference are not really distinct; in giving one, there is a tacit appeal to the other. To say they are distinct would be like saying that there are two forms of explanation of actions, one in terms of belief, one in terms of desire. 'He took his umbrella because he thought it was going to rain.' 'She went to London because she wanted to see Mike.' The former could be defeated by pointing out that he loves getting soaked by rain; the latter could be defeated by pointing out that she knew very well that Mike was in America.

So we have:

1. H. Because E. And (probably) $\neg H \Rightarrow \neg E$.

Defeated if (probably) $H \Rightarrow \neg E$.

Undefeated (so far) if (probably) $H \Rightarrow E$

2. H. Because E. And (probably) $H \Rightarrow E$.

Defeated if (probably) $\neg H \Rightarrow E$.

Undefeated (so far) if (probably) $\neg H \Rightarrow \neg E$.

That is, for a good argument from E to H, we want it to be probable that if H had not been the case, E would not have been the case; *and* we want it to be probable that if H had been the case, E would have been the case.

These facts are captured by a time-honoured principle of probabilistic reasoning, a form of Bayes's Theorem:

$$\frac{p_O(H)}{p_O(\neg H)} \times \frac{p_O(E \text{ if } H)}{p_O(E \text{ if } \neg H)} = \frac{p_O(H \text{ if } E)}{p_O(\neg H \text{ if } E)} = \frac{p_N(H)}{p_N(\neg H)}$$

The left-hand equation is a theorem of probability theory, applied to a single probability function, p_O .⁶ ('O' and 'N' stand for 'old' and 'new' respectively, and represent

⁶Note: equations need numbers. That is an idealisation, in examples like those under discussion. But it is a useful idealisation. I'm not interested in exact values, but only in orders of magnitude: 'close to 1', 'close to 0', 'around 50-50' and the like.

Note also: I have written "if" where probability theorists have "given".

The proof of the left-hand equation is as follows. $p(H\&E) = p(H) \times p(E \text{ given } H)$ [Basic Principle]. As $p(H\&E) = p(E\&H)$,

probabilities prior to learning E, and posterior to learning E.) The right-hand equation represents the recommendation that on learning E (and nothing else of relevance) your new probability for H should be your old probability for H if E (which, ceteris paribus, is reasonable--some think of this as "probabilistic modus ponens"). Eliminating the middle term, it shows how, on learning that E, your new relative values for H depend on your old together with these conditional factors. In our first example, E is "the light are off" and H is "they're at home". The inference to $\neg H$ is a good one if it's unlikely that the lights would have been off, if they were at home; unless it is also unlikely that the lights would have been off, if they were not at home.

The equation makes clear that another way the inferences may be defeated is by pointing out that the hypothesis in question *was* very unlikely, before the new evidence: "but they're always at home at this time"; or "but they promised they would be in: there must be some other explanation of the lights' being off".

Principles like the above are sometimes called principles of "updating": they tell you how to "update" your degree of belief in H, from old to new, in the light of new information E. They can, and do sometimes, have this use. But far more prevalent are instances of their use which involve "backdating" (downdating doesn't sound quite right). To use them in the up-dating way, you already have to have foreseen the possibility of the information you receive by perception or testimony; and already, before acquiring it, have

$p(H \& E)$ also equals $p(E) \times (H \text{ given } E)$. Equating the two longer expressions, we have $p(H \text{ given } E) = \frac{p(H) \times p(E \text{ given } H)}{p(E)}$. Call this equation $a=b$. Derive a similar equation for $p(\neg H \text{ given } E)$. call it $c=d$. Then the equation in the text is $\frac{a}{c} = \frac{b}{d}$. Note that $p(E)$ cancels out.

a judgement about how likely it is that you will acquire it, under various hypotheses. But we continually see, hear, read in the newspaper, etc., things which we did not anticipate the possibility of coming across. If an observation strikes you as in need of explanation, or as the possible basis of an inference relevant to your concerns, you start there, and ask yourself: how likely *was* it that I *would* get this information, if H? And, if \neg H? Your present "would haves", as I said before, record your present opinion about the acceptability of an earlier "will" (an earlier "will", incidentally, which may concern a time before you were born).⁷

Thus, we need, for the empirical inferences we make, not only judgements to the effect that such-and-such *is* (more, or less) likely; but judgements that such-and-such *was* (more, or less) likely, or likely given something else--that it was more or less likely that it *would* come about, given various hypotheses. We do not fully characterize a person's epistemic state by their present degrees of belief. One could not do much by way of empirical inference without judgements that what I am now certain does obtain, on the basis of my senses, *was* unlikely to obtain, on certain hypotheses, and likely on others. And of course, we should do our best to get such judgements right.

But if this is what we do, in explaining and drawing inferences from what we see and hear, of course we will use hindsight. My final example to illustrate this, is similar to the plane crash (and is inspired by Goldman). A long time ago, a volcano erupted. It was a slow eruption, the lava creeping onwards slowly. At that time, it was very likely that the

⁷The standard literature on probabilistic reasoning ignores the point I am stressing here. It invites the picture of reasoners as 'probabilistic machines', which attach values to all the propositions in their repertoire at all times, the values being 'up-dated' as new data is fed in. This is a wildly unrealistic picture, as well as a depressing one.

lava would eventually submerge valley *A*, but valley *B* would not be affected--given the lie of the land. However, in the unlikely event of an earthquake of a particular kind at an appropriate time, the path of the lava would very probably be switched away from valley *A*, towards valley *B*. As a matter of fact, this is what happens.

Along comes our geologist, centuries later, making his inference about the eruption. He has already found out about the earthquake. "That volcano must have erupted", he concludes, "For there is lava in valley *B* and not in valley *A*; and, given what I know about the earthquake, that is just what one would expect to find if that volcano had erupted". Also, someone who, before the eruption, said "If that volcano erupts, valley *B* will probably be submerged", was unjustified, but, in the event, right.

The point of this example is that *our inferential practices would not be well served* by rejecting counterfactuals which can only be got right with hindsight. Suppose there was a second volcano whose potential eruption, at the time in question, presented much more danger to valley *B*; but in the unlikely event of the earthquake, its lava would probably be diverted elsewhere. Only with hindsight (knowing of the earthquake) is one justified in thinking that if the second volcano had erupted, valley *B* would not have been submerged; and if the first had erupted, it would have been submerged. And it is our hindsightful judgements that stand most chance of leading us to true beliefs. This explains why our practice in evaluating these problematic counterfactuals is as it is.

Given their crucial use in empirical reasoning, then, we see why the "standard picture" was wrong. We need to take into account actual facts concerning times later than the antecedent-time. We see also, I think, why counterfactuals are best assessed probabilistically. A true/false cut-off point would not serve us well. What matters, for the

empirical inferences we make, is how likely it was that E would have happened if H, compared with how likely it was that E would have happened if $\neg H$.

We have seen a way in which our counterfactual judgements explain and justify our other beliefs. Of course they play other roles. As is implicit in several of my earlier examples, they also explain and justify our reactions of being glad or sorry, relieved or regretful, that such-and-such has happened. "I'm sorry that Fred didn't come this evening; for if he had come we would have had a fourth for Bridge". This is the retrospective version of "I want Fred to come this evening; for if he comes, we'll have a fourth for Bridge". (These cases are discussed in Adams (1997).) These positive and negative reactions to what has happened are an important part of our lives; and are assessable as reasonable or not. It's hard to believe that many of our desires, beyond the most basic hard-wired ones, would survive if we were always indifferent to what has happened. In the problem cases where the rational attitude to the forward-looking "will" differs from that of the retrospective "would", our reactions switch. I want to catch that plane. If I don't, I'll be late for the meeting. I am dismayed by missing it. On learning that the plane has crashed, my dismay switches to relief: if I had caught the plane, I wouldn't have made the meeting, or any other meetings.

I am spotted in Paris arriving, very late, for the meeting. They had just heard the news. Surprise! "She must have missed that plane", they say. "If she had caught that plane she would be dead." We'll leave open whether this is accompanied by relief or disappointment.

REFERENCES

- Adams, Ernest 1975: *The Logic of Conditionals*. Dordrecht: Reidel.
- ____1993: 'On the Rightness of Certain Counterfactuals'. *Pacific Philosophical Quarterly*, 74, pp. 1-10.
- 1998: 'Remarks on Wishes and Counterfactuals'. *Pacific Philosophical Quarterly*, 79, pp. 191-5.
- Anderson, Alan Ross 1951: 'A Note on Subjunctive and Counterfactual Conditionals'. *Analysis*, 12, pp. 35-8.
- Barker, Stephen 1998: 'Predetermination and Tense Probabilism'. *Analysis* 58, pp. 290-6.
- 1999: 'Counterfactuals, Probabilistic Counterfactuals and Causation'. *Mind* 108, pp. 427-69.
- (unpublished): 'Semifactuals, Transworld Probabilities and Causation'.
- Bennett, Jonathan 1974: Review of David Lewis, *Counterfactuals*. *Canadian Journal of Philosophy* 4, pp. 381-402.
- Edgington, Dorothy 1995: 'On Conditionals'. *Mind* 104, pp. 235-329.
- Fine, Kit 1975: Critical notice of David Lewis, *Counterfactuals*. *Mind* 84, pp. 451-8.
- Goodman, Nelson 1955: *Fact, Fiction and Forecast*. Indianapolis: Bobbs-Merrill.
- Johnson, David 1991: 'Induction and Modality'. *Philosophical Review* 100, 1991.
- Lewis, David 1973: *Counterfactuals*. Oxford: Basil Blackwell.
- 1979: 'Counterfactuals and Time's Arrow'. *Nous* 13, pp. 455-76. Reprinted with postscripts in David Lewis, *Philosophical Papers* volume 2. New York: Oxford University Press, 1986, pp. 32-66. Page references to this volume.
- Slote, Michael 1978: 'Time in Counterfactuals'. *Philosophical Review* 87, pp. 3-27.

Tichy, Pavel 1976: 'A Counterexample to the Stalnaker-Lewis Analysis of Counterfactuals'. *Philosophical Studies* 29, pp. 271-3.