Counterfactual Conundrum

JOAN WEINER

HARVARD UNIVERSITY

Counterfactual conditionals, conditionals with false antecedents which are taken to be true on other than truth functional grounds, are used extensively both in philosophical and scientific work. If we are to be justified in our claims to understanding this work we must either understand these conditionals or be able to eliminate them from these contexts. It is not clear that we can eliminate them and there seems to be some consensus that we do not fully understand them. Consequently a number of philosophers have attempted to give accounts of counterfactual conditionals. These conditionals are frequently taken to be a very special sort of abbreviated argument. In *Fact, Fiction, and Forecast*, Nelson Goodman says,

> A counterfactual is true if and only if the antecedent conjoined with relevant true statements about the attendant circumstances leads by way of a true general principle to the consequent. ([1]: 37)

Of course this is not all that clear. Goodman goes on to ask what true statements are relevant and what sort of general principles are acceptable. I will concentrate on the former question here. How can we give general criteria for what it is to be a relevant true statement for a particular counterfactual? In particular, any account of these criteria ought to be one on which unproblematic intuitively true counterfactuals are true and unproblematic intuitively false counterfactuals are false.

Goodman's discussion makes it clear that this is not an easy task at all. The initial moves are fairly easy. First we must be careful to avoid conjoining with the antecedent a statement which contradicts it. Otherwise any consequent would follow. Thus intuitively false counterfactuals would be true on an account which allows us to conjoin such statements with the antecedent. Also to avoid making intuitively true counterfac-

tuals false, we would have to rule out statments which contradict the conclusion of the counterfactual. Similarly, to avoid making counterfactuals vacuously true, we should reject statements which contradict the negation of the conclusion of the counterfactual. But once all statements which are logically incompatible with the antecedent, the consequent, and the negation of the consequent (as well as those logically incompatible given certain true lawlike generalizations) are ruled out, Goodman shows us that these criteria are not nearly restrictive enough to give us the relevant statements. The conjoined statements must not only be compatible with the antecedent, consequent, and the negation of the consequent, they must also be cotenable with the antecedent. A statement B is *cotenable* with the antecedent A if and only if it is not the case that were A true B would not be true. As before, Goodman gives us examples to show that if we do not, in our account, restrict relevant statments to those cotenable with the antecedent, intuitively false counterfactuals will be true on the account. But what kind of a statement is it to say that B is cotenable with A? It is clearly a counterfactual. Thus to explain the notion of counterfactual implication, we must use a counterfactual. However, to understand this explanation we must understand the counterfactual assertion in the explanation. Perhaps we can do this without understanding the general notion of counterfactual implication, but it seems unlikely.

After Goodman's work, there was a long period in which little work on the counterfactual problem was done. Recently, possible worlds analyses of the notion of counterfactual implication have become quite popular. The motivation for such an analysis, given that one is in the grip of the possible worlds picture, is quite simple and compelling. Goodman's discussion seems to suggest that a sentence A counterfactually implies a sentence B whenever from the two hypotheses i) that A is true, and ii) that all sentences that are actually true, except those which would not be true were A true, are true (i.e. the hypotheses that A is true and that everything is as much like it actually is as it could be, given that A is true) we can infer that B is true. A more elegant-sounding way of putting this is that in the possible world most like ours in which A is true, B is true. This seems to provide a way of spelling out the relevant true statements without making a counterfactual assertion. Robert

Stalnaker gave a possible worlds analysis of counterfactuals based on this intuition ([3]). Later, David Lewis proposed a more refined and elegant theory of this nature based on comparative similarity of possible worlds, a relation he claims is an intuitive one ([2]). Now, Goodman's suggestion that conjoined statements must be cotenable is, as he argues in *Fact, Fiction, and Forecast*, difficult to understand given that we do not already understand counterfactual implication. And it is merely a version of this analysis rephrased in terms of possible worlds on which Stalnaker and Lewis have based their analyses of counterfactual implication. If we truly and intuitively understand what it is to talk about possible worlds most like ours and a possible worlds analysis of counterfactuals is accurate both to this intuitive notion and to the intuitive notion of counterfactual implication, the problem of counterfactuals has been solved. I will discuss this question, concentrating on Lewis's account. Unfortunately, it will become clear that if the notions of counterfactual implication and comparative similarity of possible worlds interrelate as Lewis says they do, then the notion of comparative similarity of possible worlds is not intuitively well-understood at all. Rather it is an artificial relation defined in terms of certain limited syntactic properties of counterfactual implication. Hence Lewis's account suffers ultimately from the cotenability problem Goodman discusses.

Before I proceed, I would like to make a few comments about what an account of counterfactuals should be like. We seem to think true counterfactual conditionals are true because a certain relation holds between the antecedent and the consequent. We do not think counterfactual conditionals like "If I had struck this match, there would have been a famine in India" are true because there seems to be no connection between the antecedent and the consequent. Any account of counterfactuals should be an account of this relation whose holding or not determines the truth values of counterfactuals. Thus an account on which counterfactuals such as the above one come out true will be a bad account. Also, we do use counterfactual conditionals and we do have intuitions about the truth or falsity of some of them. A good account of counterfactual conditionals will preserve our intuitions about the truth values of particular counterfactuals. Throughout this discussion I will apply these minimal criteria of adequacy to various interpretations of Lewis's account.

Lewis's account is, briefly, as follows. The analysis of counterfactual conditionals is based on a notion of comparative similarity of possible worlds which is assumed to be intuitively clear. On this account, there are a number of possible worlds (or possible states of affairs) and some of these are (intuitively) more like the actual world than others. If a possible world, W, is more like the actual world than another possible world, X, then W is said to be closer to the actual world than X. (Since I will only be dealing with the truth values of counterfactuals in our world, I will sometimes say that W is closer than X.) The set of possible world can be viewed as a filled sphere with the actual world at the center and some worlds closer to the center than others. The sphere contains a system of subspheres. If W is closer to the actual world than X, then all subspheres which contain X also contain W and some subspheres contain W but not X. Lewis gives his precise explication thus: *A counterfactually implies B* if and only if either no subsphere contains A-worlds (worlds in which A is true), or in some subsphere which contains some A-worlds all A-worlds are B-worlds. In terms of the closer-than relation this amounts to saying that either there are no A-worlds or all A-worlds at least as close as some non-actual world are B-worlds.

Lewis motivates his explication by this description of the intuitive interrelation of counterfactual implication and closeness:

> 'If kanagroos 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 ([2]: 1).

Since Lewis says that counterfactual implication and closeness are well-understood, although vague, we can assume that he expects that we understand his description of a 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. I will argue that if we accept Lewis's account, we cannot have any intuitions at all about what such a state of affairs would be like. Hence we do not really understand such descriptions.

I will begin by assuming i) Lewis's explication, ii) the above intuitive relation between comparative closeness of

possible worlds and counterfactual implication, and iii) that we have a good idea of what sorts of things are true in the closest A-worlds for some false A's (or that we know something about what the A-worlds in some sufficiently small sphere look like.) What will the histories of close A-worlds be like? Since A is false, the histories of these worlds must differ at some times from that of the actual world. I will try to show that for any false A there are arbitrarily close A-worlds which differ in history arbitrarily early from our world. Consequently, the comparative similarity relation involved in counterfactual analysis must, in effect, be a relation between non-actual worlds with histories at all times different from that of the actual world. This is the sort of relation, then, which must be well-understood if Lewis's claim to have given a clear account of counterfactual implication is correct.

For each A-world either there is a time of first difference or the history must be different at all times from our history. If inside every sphere there are A-worlds with arbitrarily early differences in history from ours, I need give no argument. I will assume, therefore, that for some false sentence A there is a sphere S and a time t such that every A-world inside S differs from the actual world no earlier than at t. By a discussion of counterfactuals which present no special problems, I will be able to exhibit a general method for showing that there are arbitrarily close A-worlds with arbitrarily early differences in history. I start by giving in detail an example of the method. Suppose I have on my desk an eraser and a standing mirror.[1] In the actual world the eraser (which is in position p) is not reflected in the mirror. Let A be: the eraser is in position q (one foot to the left of position p). One might suppose that the closest A-worlds to ours will differ first at some time t at which the eraser is miraculously at q rather than p. Since Lewis seems to think that differences before the time at which the event described by the antecedent occurs are of great importance and that enough such difference might outweigh the difference involved in a minor miracle, we may be inclined to think that the closest worlds will be those with the latest first difference. Clearly the latest possible first difference between A-worlds and our world will occur when the event described by the antecedent occurs. But we will soon see that this time of first difference will not do. Consider the following intuitively true counterfactual:

> If the eraser were in position q, it would be reflected in the mirror.

How should this counterfactual be evaluated in terms of the closer than relation? We have agreed, for our first approximation, that the closest worlds (according to the similarity relation weighted appropriately for purposes of finding the truth value of this counterfactual) are those which are indistinguishable from the actual world until that time t at which the eraser is miraculously in position q. Let r be the amount of time it would take light to travel from the eraser in position q to the mirror. Then at all times before $t + r$ the eraser is not reflected in the mirror. Thus at all the closest possible worlds, the eraser is not reflected at t. However, if we were to fill in the implicit time parameters of our original counterfactual, the result would be:

> If the eraser were in position q at t, it would be reflected in the mirror at t.

Thus we cannot consider, for purposes of evaluating this counterfactual, only those worlds which differ first when the event described by the antecedent occurs. If we were to do this, many intuitively true counterfactuals would come out false under the analysis.

Perhaps this shows only that we cannot calculate the time of first difference from the given sentence. This, in itself, is not damning. It is similarly easy to see that we cannot calculate the time of first difference from most counterfactuals, even if they contain explicit time parameters. To defend Lewis's analysis against the claim that the eraser case has produced a counter-intuitive result, we might simply say that the choice of t was wrong. In fact, it is true that the choice of t was wrong. But I will try to establish that no choice of t can be right. To support this claim, let us consider in addition the following situation. Suppose I could have applied for a grant a year ago and, had I applied and my application been approved, I would have received my first stipend check today. The counterfactual:

> If I had received my first stipend check today, my name would be known to the committee awarding the grant.

is intuitively true. It does not seem unreasonable to assume that all the closest worlds in which I received my first stipend check today are worlds in which I applied for the grant a year ago. I did not apply for it a year ago, thus all sufficiently close worlds in which I received the check differ at least a year ago. Consequently, I claim, there must be eraser-at-q-worlds (A-worlds) which differ at least a year ago. If not, then all closest A-worlds are worlds in which I did not receive a stipend check, i.e. the counterfactual:

> If the eraser were in position q, I would not have received a stipend check today

is true under the Lewis account. Do we want such results? No. There seems to be no connection between the antecedent and the consequent. Clearly such counterfactuals are not intuitively true. Any account which has this result fails to satisfy the minimal criteria for adequacy set down earlier. This completes the discussion of an example of the method for pushing the required time of first difference back arbitrarily far. The method turns on the fact that our intuitions concerning times of first difference are a result of our beliefs about the causes of certain events and what times these causes could latest have occurred. Thus to find a close A-world which differs earlier than at some time, t, we need first to find a false sentence B which is intuitively unrelated to A and which, in all reasonably close B-worlds, is the report of a result of an event which took place before t. Second choose another effect of the same event which is described by some C such that B intuitively counterfactually implies C. If we can always find such events, it is easily seen that given false A, sphere S, and time t, there is a possible A-world inside S which differs from the actual world before t.

Suppose, however, we want to accept Lewis's analysis but object to my claim that inside every sphere there are worlds with arbitrarily early differences. One might say that there are times sufficiently far back that we simply can say nothing about the history of our world then. In other words, we might say that we cannot find the required examples to push the time of first difference arbitrarily far back, since there are times before which we know nothing about what events did or did not occur. Then the time of first difference might be any time before such a time. This seems to be a somewhat feeble

argument. There does not seem to be that much difference between the world's history at times about which we can talk knowledgeably and the world's history at earlier times. So why should we assume that there is *sufficient* difference to merit the claim that there are earlier times at which there could have been points of first difference in all sufficiently close possible worlds? (In fact, I have argued that as long as we can talk knowledgeably about the history of the actual world at a particular time, points of first difference from the actual world must be at least that early. Thus if we believe there are points of first difference, we ought to believe that there are times about whose histories we can, in principle, know nothing.)

There is another objection to my claim that if Lewis's account is correct, we must have intuitions about comparative similarity of worlds which differ arbitrarily early from ours. This objection has to do with the miracle hypothesis. Why not assume that there can be a point of first difference, but not a point of first difference at which a miracle occurs? This seems to stem from the following sorts of intuitions. It seems that counterfactuals like "If the eraser were in position q, I would have moved it there," "If I had moved the eraser to position q, it would have been in my way at an earlier time," etc. are true. We can argue in two directions from this. One direction will clearly take us to my original conclusion that we can push time differences arbitrarily far back. In a second direction, we could argue that points of first difference can be traced back to random or undetermined events. To see the difficulty with arguing this way, it should be noted that the sort of claim which must be made is a very strong one. These events, for instance, must not only be random, there also must not be any necessary preconditions for their occurrences. (If there were, we could generate backwards-directed counterfactuals which would show that there must be earlier differences.) Such typical examples of undetermined events as dice throwing or people's decisions for this reason will not suffice. That is, the randomness must be intrinsic. But even if each event could be traced back to an intrinsically random one, we would have to be able to show that the set of random events to which all events are traceable does not contain arbitrarily early events. It does not seem possible to make out such an argument.

I will assume, from this point on, that for each counterfactual there are relevant sufficiently similar non-actual

worlds with histories different at arbitrarily early times from that of the actual world. Thus general statements about what very close A-worlds are like must amount to statements about A-worlds with histories which differ at all times from ours. The important question, then, is whether or not there is a well-understood intuitive notion of comparative similarity of possible worlds with histories which differ at all times from ours. Even if Lewis is correct in his claim that we have an intuitive notion of comparative similarity of possible worlds, it is not obvious that this notion will apply to the counterfactual analysis. How do we talk about comparing worlds with arbitrarily early differences in histories? In particular, how do we know what sort of things would happen in worlds with distinct histories arbitrarily far back? Lewis likens comparative similarity of possible worlds to comparative similarity of cities, but there is an important difference here. While cities may have ill-defined boundaries and while there may be many important facts about them which we cannot know, we do know many things about cities. This is not true of possible worlds with histories which differ from ours at all times. For how are we to decide what other things are true at close A-worlds where A is false? Suppose we want to argue that some sentence B is true in arbitrarily close A-worlds. (This is the weakest sort of claim required in applications of Lewis's analysis). We know that there are very strange AB-worlds, how do we know there are any close ones? When A describes an event, we are likely to do what was done in the eraser case, i.e. construct a likely chain of events leading up to the occurrence described by A and argue that had this chain occurred, the event described by B would. There will probably be several of these chains each of which, we would claim, describes part of a nearby possible A-world. Our motivation for describing such chains would have been a belief that, had A occurred, it would have been a result of the occurrences in one of the chains. But how are we to justify, for any of these chains, the claim that the events in this chain occur in nearby A-worlds? Again we would argue by exhibiting likely chains which reach even farther back. If we cannot continue the process indefinitely, we cannot justify by this procedure the claim that the longest chain formed occurs in any nearby A-world.

Either we are able to continue such a chain indefinitely or we must stop at some point in time at which our intuitions are

of little help. If we must stop the procedure, then we must admit that we do not know what nearby A-worlds are like. We have no way of justifying the claim that for an arbitrarily close A-world *this* time at which we have stopped is a time of first difference, and we have no way of knowing what consequences changes before this time will have later on. Consequently we have no way of knowing what nearby A-worlds are like at t. On the other hand, perhaps we can argue that such a chain could be continued indefinitely. But if the procedure never stops, it will never provide us with answers to our questions about A-worlds. Thus we wll need another sort of argument to show that there are arbitrarily close A-worlds which are B-worlds.

There is, given Lewis's analysis, one effective way of doing this. We can use the analysis and intuitively true counterfactuals to show what must be true in sufficiently close worlds. This is the method which was used earlier in the paper to push times of first difference back. Application of the method allows us to say any number of things about sufficiently close worlds. But then how do we know that *this* relation of 'comparative closeness' is an intuitive one? We cannot justify by intuitive facts about the comparative similarity relation plus facts about the world, any of the claims about comparative similarity of possible worlds which we must make to apply Lewis's analysis. We can justify these claims about comparative similarity only in terms of Lewis's analysis. Thus the analysis seems to be an analysis of an artificial relation in terms of the natural notion of counterfactual implication. In other words, we have no way of justifying the claims that Lewis's analysis holds and the comparative similarity relation involved is a natural one. Furthermore, that we can only justify claims about comparative similarity by discussion of intuitively true counterfactuals shows that Lewis's account is vulnerable to Goodman's cotenability problem. Goodman has shown that in order to give a certain sort of account of counterfactual conditionals, one must make a counterfactual assertion. I have tried to show that if Lewis's analysis is correct, any understanding of what is true at nearby possible worlds is based on assumptions about the truth of certain counterfactuals. Thus possible worlds accounts of counterfactuals have no advantage with respect to the cotenability problem over any other accounts.

While this attack is explicitly directed only at Lewis's account of counterfactuals, the real argument has nothing to do with this particular account. The real argument is, basically, that we have no way of knowing what would be the case, were A true, since so many things would be different if A were true. This can be phrased either as a problem with possible worlds or, given an account of counterfactuals, as a problem with understanding counterfactuals.[2]

REFERENCES

[1] Nelson Goodman, *Fact, Fiction, and Forecast* (Indianapolis: Bobbs-Merrill, 1965).
[2] David Lewis, *Counterfactuals* (Oxford: Basil Blackwell, 1973).
[3] Robert Stalnaker, "A Theory of Conditionals," in *Studies in Logical Theory*, ed. by N. Rescher (Oxford: Blackwell, 1968).

NOTES

[1]This example is due to Hilary Putnam.

[2]I would like to thank Burton Dreben, David Hills, Mark Kaplan, Hilary Putnam, and Thomas Ricketts for reading and criticizing drafts of this paper, and G. Lee Bowie for conversations and unpublished work.