The Similarity Approach to Counterfactuals: Some Problems

G. LEE BOWIE

MT. HOLYOKE COLLEGE

Although modal logicians have traditionally been interested in counterfactuals, only recently has the application of possible world semantics to the problem of counterfactuals begun to yield formal semantics which appear to capture the truth conditions of ordinary English counterfactuals.[1] The advance has been brought about by defining similarity structures on sets of possible worlds. This use of similarity among possible worlds in the analysis of counterfactuals has been the subject of considerable theoretical development. Through the work of Lewis, Stalnaker, van Fraassen, Thomason, *et. al.*[2], the basic idea has been modified and generalized in the direction of higher order quantification (over modalities), various indexings for selection functions, impossible worlds as limits of sequences of possible worlds, counternecessaries, countercomparatives, counterfactual probabilities, and so on, all accompanied by elegant model-theoretic notions and axiomatics. Our natural tendency to preserve the products of great labor is, in this case, reinforced by the power and elegance of the theories themselves. But however elegant and powerful the similarity approach is in its full theoretical development, there are difficulties at its foundations—difficulties which cast doubt on the use of the notion of comparative similarity of possible worlds in the analysis of counterfactuals.

This paper, then, is concerned to point out some of these difficulties, which surround the basic assumptions of the similarity approach in its application to the problem of counterfactuals. I shall not be arguing that the similarity approach does not capture the logic of counterfactual inference. In fact it appears that the similarity approach, as developed by Lewis, *does* capture the logic of counterfactual inference. Indeed, the fact that it does is a very strong argument on its behalf. Nor shall I be arguing that the notion of similarity of worlds is vague and imprecise (although, to be sure, it *is* vague and imprecise). What I shall be arguing is that no plausible understanding of the notion of similarity generates a correct analysis of the truth conditions of ordinary counterfactuals. More formally, while theories based on similarity can verify as valid all and only those ordinary counterfactuals which seem, in fact, to be valid, nonetheless no plausible understanding of the notion of similarity will give us as true (in the intended model) just those ordinary counterfactuals which seem, in fact, to be true. Thus while the similarity approach generates a successful logic, it does not generate a successful analysis.

Before proceeding to similarity theories, it will be useful to review the classical approach to counterfactuals. This will both highlight the successes of the similarity approach, and enable us to connect the failures of the similarity approach with the failures of the classical approach. I follow Lewis in using '$A \Box\!\!\rightarrow B$' to represent the counterfactual 'if A were the case then B would be the case'. In addition, '$\rightarrow$' is the material conditional, $\bar{A}$ is the negation of A, and an A-world is a possible world in which A is true. Now the simplest sort of possible worlds analysis of counterfactuals would be this: for each world *i* there is a class of worlds C_i such that a counterfactual $A \Box\!\!\rightarrow B$ is true in *i* iff $A \rightarrow B$ is true in every world in C_i. Lewis calls this theory a strict conditional theory, since on it we can understand $A \Box\!\!\rightarrow B$ as $\Box(A \rightarrow B)$, where C_i is then just the set of worlds accessible from *i* with respect to the modality $\Box$. In the case where C_i is just the set of worlds at which some formula S_i (whose choice may depend on *i*) is true, we may restate the theory as follows: a counterfactual $A \Box\!\!\rightarrow B$ is true at *i* iff $A \rightarrow B$ is true at every world (accessible from *i*) at which S_i is true [iff $(S_i \ \& \ A) \rightarrow B$ is necessary].

It is clear that there is no choice of S_i (or of C_i) which captures the counterfactual conditional. For choose A as any proposition incompatible with S_i. Then no matter what B is, it will (vacuously) be true in every world in which $(S_i \ \& \ A)$ is true. Hence on this theory $A \Box\!\!\rightarrow B$ will be vacuously true whenever A is incompatible with S_i. More generally, if B is true at every world at which $(S_i \ \& \ A)$ is true, then *a fortiori* B is true at every world at which (S_i & A & C) is true. Hence on a strict conditional theory $A \square\!\!\rightarrow B$ entails $(A \; \& \; C) \square\!\!\rightarrow B$.

But this inference is not sound for counterfactuals. For example, if Mort were to go to the store tomorrow he would buy bananas. It does not follow that if Mort were to go to the store tomorrow and get shot on his way to the checkout stand, he would buy bananas.

The arguments above show that there is no set of worlds such that for every A and B, $A \square\!\!\rightarrow B$ is true iff $A \rightarrow B$ is true at every world in that set. The source of the difficulty is clear. When we ask whether $A \square\!\!\rightarrow B$, we are asking, it would seem, about the truth value of B in relatively familiar worlds in which A is true. Interpreting the counterfactual as a strict conditional requires that we decide on a standard of familiarity before we know what A is. S encodes the standard of familiarity. But having chosen such a standard, we may then wish to ask what would happen if that standard were not met. But we cannot do this. The standard has already been fixed. Thus the first objection above. Similarly, having chosen some familiar set of worlds to look at in which Mort goes to the store, having chosen to ignore as too bizarre worlds in which he gets shot while shopping, we may turn around and ask what would happen if Mort were to go to the store and get shot. Now we go wrong if we ignore worlds in which Mort gets shot, whereas before we would go wrong if we were to consider such worlds. Thus the second objection above.

Given a possible worlds approach to counterfactuals, it is clear that in evaluating the truth of $A \square\!\!\rightarrow B$ we want to look at the truth of $A \rightarrow B$ in some set of worlds. From the discussion above it is clear that the choice of the set of worlds must be variable, dependent on A, and perhaps on B.

The classical approach to counterfactuals takes this into account. According to the classical account, a counterfactual $A \square\!\!\rightarrow B$ is true iff there is some true S such that B is true in every world in which (S & A) is true. In contrast to the strict conditional theory, the choice of S may depend on A (or on B). S in general will be a huge conjunction which will, intuitively, tell us enough about what our world is like other than the fact (if it is a fact) that $\bar{A}$, so that when we conjoin it with A, B is entailed. For example, if we are wondering whether Mort would buy bananas if he were to go to the store tomorrow, S would contain natural laws, enough information to determine that the store will have bananas, to determine Mort's proclivities toward banana-buying, as well as information about earthquakes, homocidal maniacs and other factors which would, if they obtained, bar Mort's banana-buying proclivities from their natural fruition.

We have required that S be true, but clearly more is required. For whenever $A \rightarrow B$ is true, it will, together with A, entail B. (When $A \rightarrow B$ is false, so is $A \Box\!\!\rightarrow B$.) So if we can always choose $A \rightarrow B$ as S, the counterfactual conditional will collapse into the material conditional. Goodman terms 'co-tenable' those truths S which are appropriate for conjoining with A when entertaining A as a counterfactual hypothesis. If we are wondering what would happen if Mort were to go to the store tomorrow, then truths of the sort noted earlier would be co-tenable with the proposition that Mort goes to the store tomorrow.

Now, why is $A \rightarrow B$ (when true) not always co-tenable with A? At first glance, this might seem to be another of those cases in which we would like to be able to draw a distinction between "real" facts (like natural laws, the fact that there will be no homocidal maniacs in the store tomorrow, etc.) and "phoney" facts (like the fact that $A \rightarrow B$). Having drawn such a distinction, we would then say that a truth is co-tenable with A just in case it represented a real fact and it was (in some appropriate sense) compatible with A.

Now quite apart from the fact that attempts to so distinguish the real facts from the phoney ones have not been notable successes, the attempt would be misguided in this context. Martha's salary is $100,000, and her property taxes are $5,000. Out of an equal desire to donate 10% of her salary annually and to donate twice her property tax annually, she gives $10,000 every year to the Flat Earth Society. Now, what would happen if Martha were president (of the U.S.). The fact that Martha always gives 10% of her salary and the fact that the president's salary is $200,000 entail (together with the hypothesis that Martha is president) that she will give $20,000. But the fact that she always gives twice her property tax and the fact that the president pays no property tax on the White House entail (together with the hypothesis that Martha is president) that she will give nothing. Yet all of the facts alluded to seem (to me) to be as intuitively "real" as facts which typically support counterfactuals. There are, it would seem, too many "real" facts around, so we must make some further choice. Unfortunately, among the facts relied on above, the ones that we want are just those facts that *would remain* facts if Martha were president. So we have come full circle, since in order to decide what would be the case if A were the case we must find those truths which are co-tenable with A. And to do that, we must find which facts are "real" facts. But among those, the only ones in which we are interested are those which would remain facts if A were the case, so we must again know what would be the case if A were the case.

If the truths co-tenable with A are just those which intuitively "support" the counterfactual A□→B then there is good reason to believe that in general the truths co-tenable with A will be just those truths which would remain truths if A were the case. For the claim that S supports the counterfactual A□→B can always be defeated by showing it to be false that if A were the case then S would be the case (as reflection on a few examples shows). So only truths which would remain truths if A were the case support A□→B. And while some proper subsets of those truths would *suffice* to support all true counterfactuals with antecedent A (for example, the set of all truths A→B which would remain truths if A were the case), examination of cases makes it appear likely that *any* truth which would remain true if A were the case can be enlisted in support of the counterfactual A□→B (with the proviso, perhaps, that it does not itself entail B). A solid argument is not possible here, due to the fact that the notion of "support" appealed to remains a rough, intuitive notion. The considerations raised above, however, make it appear plausible that it will be no easier to characterize carefully the truths which support the counterfactual A□→B than it will be to characterize the propositions which are counterfactually implied by A and to select from them those which are true (and, perhaps, which do not imply B). Thus if the truths co-tenable with A are just those which support counterfactuals with antecedent A, it appears that in order to give an account of co-tenability, the classical theory will have to give an account of the counterfactual conditional itself.[3]

David Lewis provides an elegant account of counterfactuals which, to be understood as an analysis of the counterfactual conditional, requires (in addition to the semantics of possible worlds) only the intuitive notion of comparative over-all similarity taken as a 3-place relation among possible worlds. On Lewis' account, given a possible world i, there is a set of spheres associated with i. The set is nested, and closed under union and intersection. Each sphere is a set of possible worlds. Intuitively, if there is a sphere associated with i such that j is inside it and k is outside it, then j is more similar to i than k is. We need only consider the usual case where i is a member of every non-empty sphere associated with itself. Thus we will speak of the spheres associated with i as spheres *about* i. A sphere is A-permitting iff there is some world in it at which A is true. Then $A \mathrel{\Box\!\!\rightarrow} B$ is true at i iff either there is no A-permitting sphere about i, or there is some A-permitting sphere about i such that $A \rightarrow B$ is true at every world in that sphere. Phrased in terms of similarity, we get the following for the non-vacuous case: $A \mathrel{\Box\!\!\rightarrow} B$ is true at i if there is a world at which A is true such that at no world at least as similar to i is A true and B false.

The characterization in terms of spheres and the characterization in terms of the 3-place similarity relation (j is more similar to i than k is) are equivalent. Given a system of spheres, we can say that j is more similar to i than k is iff for some sphere, S, about i, j is inside S and k is not. Similarly, given the comparative similarity relation, we can define a system of spheres as follows. For each world i, S is a sphere about i iff every world in S is more similar to i than any world not in S.[4] There is then a natural isomorphism between sphere-systems and similarity systems which preserves truth value of all formulas at all worlds. I have described Lewis' semantics initially in terms of spheres, since the truth definition is simpler and since it is Lewis' preferred formulation. I will rely on the similarity formulation in the sequel, since it is a more convenient vehicle for exposing the difficulties which I raise. Naturally, in virtue of the close correspondence between sphere structures and similarity structures, any doubts about similarity generates corresponding doubts about spheres; so no purpose other than ease of exposition is served by the decision to focus on the similarity characterization. Finally although the accessibility relation enters into Lewis' account, it is not relevant to my discussion, and I have ignored it. (Ignoring it is equivalent to supposing it to hold between all possible worlds.)

If we assume for any world *i* and proposition A, that if there is any world at which A is true then there is a closest world at which A is true (that is a world at which A is true which is more similar to *i* than any other world at which A is true), we get as a special case Stalnaker's theory: That $A \Box\!\!\rightarrow B$ is true at *i* iff B is true at the closest (most similar) world to *i* at which A is true.

Lewis' theory seems superior to Stalnaker's. It is more general, since Stalnaker's theory can be obtained as a special case of Lewis' by placing suitable restrictions on the system of spheres (or on the similarity relation). Stalnaker's models do not seem to be generated by our intuitive notion of similarity. For example suppose, contrary to fact, that Mort were standing more than 120 feet away from the Washington Monument. It would seem that given any world in which Mort were standing $120 + \delta$ feet away, there is a world more similar to ours in which Mort is standing only $120 + \delta/2$ feet away. Thus there seems to be no closest world in which Mort is standing more than 120 feet away from the Washington Monument. Furthermore, it would seem as though two different worlds could be equally similar to a given world. Consider a world in which three pool balls are lined up on a pool table, the 8-ball between the 7-ball and the 9-ball. Is a world in which the 8-ball is 3 inches closer to the 7-ball more or less like the given world than one in which the 8-ball is 3 inches closer to the 9-ball? It would seem as though the two worlds are equally similar to the given world. But this cannot be so in Stalnaker's models. Thus Lewis' models better capture our intuitive notion of similarity.

The more telling consideration is this: Stalnaker's theory validates the scheme $(A \Box\!\!\rightarrow B) \vee (A \Box\!\!\rightarrow \bar{B})$. But this does not seem correct. In the example above, which of the following counterfactuals is true: (i) if the 8-ball were 3 inches closer to either ball, it would be 3 inches closer to the 7-ball, (ii) if the 8-ball were 3 inches closer to either ball it would be 3 inches closer to the 9-ball? Surely if the 8-ball were 3 inches closer to either ball then it would either be 3 inches closer to the 7-ball or it would be 3 inches closer to the 9-ball. But this does not force us to accept either (i) or (ii), and intuition seems to have it that neither (i) nor (ii) is true. But either (i) or (ii) must be true on Stalnaker's theory.

Both Stalnaker and Lewis give formal semantics and axiomatics for counterfactual logic. Structures are defined over arbitrary sets, understood as sets of possible worlds, and arbitrary orderings on those sets (or their equivalents, as in Lewis' spheres), understood as similarity orderings. Both intend their systems to generate an analysis of counterfactuals in the sense that one gets the truth conditions for ordinary counterfactuals by looking at the truth conditions of their corresponding formulas in the standard interpretation, that is, the interpretation which takes its set of possible worlds to be the real set of possible worlds, and which takes the relation on possible worlds to be the real similarity relation on the set of possible worlds. What I shall try to show in the sequel is that there is no plausible understanding of the real similarity relation which gives us the correct truth conditions for ordinary counterfactuals in either Lewis' or Stalnaker's theory. For ease of exposition it will be convenient to speak, à la Stalnaker, of the A-world which is closest to ours, or the A-world most like ours, or the A-world most similar to ours (these are interchangeable). All of my remarks can be reformulated so as to avoid assuming that there is a closest A-world, though generally at the cost of more complex sentence structure.

Now, when we compare two worlds for over-all similarity, in what respects do we compare them? It follows from Lewis' account that we must compare them in all respects. For if there were respects in which a world might differ from *i* which did not count at all in judging its overall similarity to *i*, then there could be a world which differed from *i* but only in those respects whch counted for nothing in judging overall similarity. This would entail the existence of a world which was not identical with *i*, but which was just as similar to *i* as *i* is to itself. But Lewis adopts the "centering assumption," which requires that no world be as similar to *i* as *i* is to itself. Furthermore, Lewis evidently intends that worlds be compared in their temporal totality. That is, the basis of comparison is the entire world, in its spatial and temporal totality, and that totality is to be compared with the spatial and temporal totality of other worlds in every possible respect of comparison. Although some respects of comparison may count for little in judging overall similarity, no respects of comparison can count for nothing.

But this method of comparison does not seem to give us correct truth conditions. Consider: I am standing with my finger on the ultimate doomsday button. The button and associated mechanisms have been thoroughly tested, and are reliable. If I were to push it, the entire universe would explode—*unless* a loose piece of metal drops down as I push the button in such a way as to jam the button. But it won't. It is loose, but not loose enough. Yes, I see that if I pushed the button, the universe would explode. But what happens in the world most like ours in which I push the button? There can be only one answer—the piece of metal falls anyway and jams the button. For surely a world in which the button gets jammed is far more like ours than a world in which it doesn't and the entire universe explodes. Even if the piece of metal had to come from all the way across the room, or if the button had to disintegrate, the resulting world would be more like ours than a world which explodes. In fact we can say with some confidence that in the world most like ours where I push the button, the nearest relatively unattached small object flies under the button and jams it, for surely a temporary local breakdown in the laws of mechanics would preserve similarity far more than world cataclysm. So, the example continues, if I push the button, the world would not explode. But we know that if I push the button, the world will explode. We can inspect the mechanism, do simulated trial runs, etc. So the theory is wrong. It is wrong because it makes change too difficult. If there is some simple way to keep the world on track, it will happen. If things can go right, they will.

If the objection stands, the theories we are considering would be clearly unacceptable, and the similarity approach would have to go. Now it might be thought that the objection can be met by placing more weight on laws of nature in determining similarity. This approach fails for several reasons. First, no matter how much weight we place on laws, we must still place some weight on particular matters of fact. To do otherwise would be to do injustice to more mundane counterfactuals and to run afoul of the centering principle. But whatever reasonable distribution of weights we decide upon, by making the results of preserving the laws sufficiently spectacular, and by making the breakdown in laws sufficiently unexciting (as in the example), we will find cases where a small, local violation of the laws preserves similarity more than strict adherence to the laws.

Secondly, in entertaining a counterfactual with false antecedent, the only worlds in which the antecedent is true but where the laws of this world are preserved might be worlds which are unacceptably different from ours. As Lewis[5] says,

> Suppose that the laws prevailing at a world *i* are deterministic, as we used to think the laws of our own world were. Suppose a certain roulette wheel in this deterministic world *i* stops on black at a time *t*, and consider the counterfactual antecedent that it stopped on red. What sort of antecedent worlds are closet to *i*? On the one hand, we have antecedent worlds where the deterministic laws of *i* hold without exception, but where the wheel is determined to stop on red by particular facts different from those of *i*. Since the laws are deterministic, the particular facts must be different at all times before *t*, no matter how far back.
>
> ... On the other hand, we have antecedent-worlds that are exactly like *i* until *t* or shortly before; where the laws of *i* hold *almost* without exception; but where a small, localized, inconspicuous miracle at *t* or just before permits the wheel to stop on red in violation of the laws. Laws are very important, but great masses of particular fact count for something too; and a localized violation is not the most serious sort of difference of law.

Thirdly, in testing scientific hypotheses we must take some of the laws rather lightly. Suppose for example that we are testing part of the special theory of relativity for the first time, by checking on the apparent position of Mercury during a solar eclipse. Classical mechanics has it that Mercury will appear to be where it is, at point C. Special relativity has it that Mercury will appear to be at point S, distinct from C. Interestingly, Mercury appears to be at S, and we infer that classical mechanics was wrong. But let us entertain the counterfactual hypothesis that Mercury had appeared to be at point C. There are two hypotheses (at least): H1-that although the photons were disturbed by the mass of the sun, random displacement of the photons compensated for this disturbance, making Mercury appear at point C. On H1, special relativity was right even though Mercury appeared to be at C, because there were compensating random influences. H2 has it that the mass of the sun did not disturb the passage of the photons, and classical mechanics was right. Now the point is that in the actual case, we clearly reject H1 as an explanation of our hypothesized failure to confirm special relativity, and assert instead that if Mercury had appeared at point C, then relativistic mechanics would have been incorrect. Our acceptance of H2 in this counterfactual situation requires that we count as closer a world in which relativistic mechanics is incorrect, a world in which the laws of nature are violated, than a world which is fully accounted for without violation of any laws of nature, but which is so improbable as to be disregarded. We conclude with Lewis that "the preeminence of laws of nature among cotenable factual premises is only a matter of degree."[6]

Fortunately, there is an easier way to meet the objection. It can be met by making clear that the world we are comparing with ours is not being compared in virtue of its temporal totality. We must require only that its history up to (and perhaps including) the time at which A is true (for counterfactual A□→B) is most like the history of this world. In the example, we are to imagine standing in the room, finger on the button; the stage is set, and everything so far is as much as possible like things are here. At this point—I have just pushed the button—we stop worrying about how close the worlds are; we just sit back to wait and see what happens. In this case we needn't wait long—the world explodes. It no longer matters, though, that the world has just become very different from ours. We just stopped comparing. We made the worlds as close as possible up to and including the time of the button-push, and then let nature take its course. Since that world was like ours, the metal was loose but not too loose, the mechanism worked, etc. And because all that was the case, the world blew up, just as this world would have if I had pushed the button. Once the basis of comparison is made clear, the counterexample is blocked. Unfortunately, the solution raises new difficulties.

Paradigm counterfactuals are forward-directed. They conjecture what would happen in the future if things were different in the past or present, or what would happen in the future if things were different at the same time or earlier in the future. But counterfactuals can also be past-directed.

There is an alarming instability in the structure of the universe which makes it liable to cataclysm. Because we find it comforting, we have invented an early warning device (EWD) which sends out a signal in advance of the universal destruction. The device, for proper operation, is situated at the center of mass of the universe. There are two societies—one, Society Near, so close to the center of the universe that if it receives a signal, there will be 3 years before the cataclysm begins. The other, Society Far, is so far from the center of the universe that by the time it receives the warning signal, ⅔ of the universe will already have been destroyed, and the remaining ⅓ will go soon. Now, entertain the two counterfactual hypotheses: 'Society Near receives a signal from the EWD' and 'Society Far receives a signal from the EWD'. In the first case, since the device has been tested and is highly reliable, all works well. Since we are only comparing the histories of the universes up to the time when Society Near receives the signal, a universe in which the signal is received and is genuine will have a history more like that of our universe than a universe in which a signal is received through a malfunction of this highly reliable device. We may say, then, that if Society Near were to receive a signal, the cataclysm would begin 3 years thence.

On the other hand, suppose Society Far receives a signal (even the same signal). *Now*, any universe in which a signal is received by Society Far, and is genuine, is a universe which has already been ⅔ destroyed. Since we are comparing the histories of such universes to the history of our universe, such a universe, being ⅔ destroyed, would be grossly different from our intact one. We are in the same position as before with the doomsday button. For any universe in which Society Far receives a signal which is genuine (hence a universe which is already ⅔ destroyed) there is a universe whose history is more similar to the history of our intact universe, in which the signal resulted from a malfunction. Hence the counterfactual 'if Society Far receives a signal, the cataclysm has begun' is false, on the similarity theory, even if we suppose Society Far to receive the *same* signal which we suppose Society Near to have received, from the same device.

So in choosing histories as our basis of comparison, we are still comparing too much. Perhaps what we are to do is to compare some world with ours only with respect to its temporal slice at the time at which A is true, ignoring, for purposes of comparion, not only its future from that time (as we did above), but also its past before that time. Thus, for the counterfactual $A \Box\!\!\rightarrow B$, we pick the A-world which is such that its temporal slice is most like ours, and see what will (did) happen in that world. Note first that unless we assume a strong form of determinism, this suggestion will require that we abandon the centering assumption. For the centering assumption requires that no world is as similar to *i* as *i* is to itself. But on the current suggestion we are comparing only temporal slices of worlds;

and unless some strong form of determinism holds, two distinct worlds will have identical slices at a given time, differing only at some earlier or later time.[7] Since we are comparing only temporal slices, two such worlds will be as similar to each other as each is to itself. But this is not permitted by the centering principle. However, since we could adopt what Lewis calls "the weak centering principle,"[8] which requires only that no world be more similar to *i* than *i* is to itself, no intractable problem is created for the similarity approach.

When we compare an A-slice of another world with our world, what are we comparing it with? When we were comparing temporally complete worlds, this was no problem. But now that we are looking only at an A-slice of a possible world, we must presumably compare it for similarity with a *slice* of our world; and the question is, which slice? Evidently we do not compare the A-slice with a current slice of this world. Suppose that it is December, and my foot is in a cast which will come off next week. If we are entertaining the counterfactual 'if I were to go skiing in February, I would need a new pair of ski boots', we would go wrong if we compared slices in which I go skiing in February with current slices, since in the slice most like the current slice of our world, I have a cast on my foot. Hence we would be asking what will happen in a world in which I go skiing in February with a cast on my foot, and this is clearly not going to give us the right answer.

Do we compare A-slices with a slice of our world which is, as it were, cotemporaneous with the A-slice?[9] That is, in the example above, do we compare A-slices with a slice of this world taken in February? Suppose, as a matter of fact, that in January I decide not to go skiing, but to cut off the left half of my foot instead. Thus in February I have only half a foot. We would go wrong if we were to compare A-slices with a slice of this world taken in February, for in that slice I have only half a foot (and would need new ski boots). But in the real counterfactual situation, if I had gone skiing in February, I would not have decided to mutilate my foot in January, and would not have needed new boots.

It is not as easy as it should have been to see where we went wrong. For in deciding that we had to compare slices, and not worlds, we tacitly assumed that we would compare A-slices with slices of our world. But in the skiing examples, the intent is clearly that among all of the A-worlds, we chose the one whose current slice is most like the current slice of our world, and then see what happens there. Thus in the first skiing example above, look among the worlds in which I go skiing in February for that world whose December is most like December in this world, and then see whether I need new boots in that world. In that world I do not have a cast on my foot in February, and I do not need new boots. The same procedure applied to the second example gives us a world in which I go skiing in February, do not mutilate my foot, and do not need new boots. Now that the basis for comparison has been made clear, things seem to work right.

Unfortunately, we have still not got it right. Return to the example of the early warning device. On the present suggestion, in evaluating the counterfactual 'if Society Far receives a signal, the cataclysm has begun', we are to look among all of those worlds in which Society Far receives a signal for the one whose current slice resembles our world's current slice. But the current slice most of any world in which the signal is genuine will be ⅔ destroyed, and will be less like our current slice than one in which the signal resulted from malfunction. Thus the counterfactual still turns out false on the similarity theory, when it should turn out to be true.

Two more examples will help illuminate the problem. Consider the false counterfactual 'if the Norman Conquest had failed, the world would be pretty much the way it is now'. The present suggestion has us look among all of the worlds in which the Norman Conquest fails for the one(s) whose current slice is most like our current slice, that is, for the one which is most like ours now. But given the possibility of compensating differences which counteract the failure of the Norman Conquest, there is at least one world in which the Conquest fails which today is pretty much like ours. Hence the world in which the Conquest fails which today is most like our world today, will be a world which is pretty much like our world today. Thus on the present suggestion, the counterfactual 'if the Norman Conquest had failed, the world would be pretty much the way it is now' comes out true. To get the right answer in this example we have to compare A-slices (slices in which the Norman Conquest has just failed) with co-temporaneous slices of our world, or alternatively, compare histories up to the time of the failure of the Norman Conquest with co-temporaneous histories of our world. While these procedures give us the right answer in this case, they failed in earlier cases.

Now consider one last example. Mort is sitting at the dinner table with a dish of magic yogurt in front of him. Mort is 6 feet tall. The yogurt works as follows: if he eats exactly one teaspoon or less, he will remain 6 feet tall. As he eats more than one teaspoon, he will grow as a linear function of the amount, until, if he eats it all, he will grow to 7 feet tall. As a matter of fact, Mort abominates yogurt, magic or not, and is not going to eat any of it, although he may dip his spoon in it and play with it. Now, consider the counterfactual 'if Mort were to eat some of this yogurt, he would remain 6 feet tall'. My intuition is that neither this counterfactual, nor the associated counterfactual, 'if Mort were to eat some of this yogurt, he would grow', is true.[10] The antecedent can be satisfied in too many ways, and there is no way of projecting which way it *would* be satisfied if it were satisfied.

But the similarity approach says that if Mort were to eat some of the yogurt, he would remain 6 feet tall. For a world in which *ceteris paribus* he is eating only a tiny bit of yogurt is more similar to our world, in which he eats no yogurt, than any world in which he eats more. Hence in the closest world(s), Mort eats only a tiny bit of yogurt and remains the same height.

It might be thought that there will be features of the situation which would give us a better answer. If, for example, Mort were an inveterate glutton, given to extremes, then perhaps we could say that he would either eat none or eat it all. If, on the other hand, he were a man of delicate sensibilities, given to caution in new ventures, then we could say that if he were to try some, he would try only a tiny bit, and would remain the same height. This would amount to saying that there are features of the situation (including Mort's character) which are more important in determining similarity, than the mere quantity of yogurt which Mort eats.

While such considerations might settle the issue more favorably in this case, the general problem remains. For however we construe similarity, if in a counterfactual A□→B, the antecedent is false but can be satisfied in a number of ways, and if there is no way of projecting which way it would be satisfied, if it were satisfied, then the similarity approach requires that it be satisfied in whatever way is closest to its not being satisfied at all. So for any antecedent A, we have the following counterfactual true: if A were the case, then A would barely be the case. Thus, if Mort weighed between 300 and 400 pounds, he would weigh between 300 and 310, between 300 and 301, etc. Here there is nothing in Mort's character or constitution which suggests any particular range within which his weight would fall if he were to weigh between 300 and 400 (except, of course, the range 300-400). We have no reasonable way of saying, for example, whether he would weigh over 350 or under 350. But the similarity approach requires that he weigh between 300 and 301 if he were to weigh between 300 and 400. This is because worlds in which he weighs between 300 and 301 are, *ceteris paribus*, more like our world (in which Mort weighs, let us say, 180) than are worlds in which he weighs over 301.[11]

Let us review our results thus far. In evaluating the truth value of the counterfactual $A \Box\!\!\rightarrow B$ at a possible world i (we suppose here that A is false in i), the similarity approach has us look at worlds j, in which A is true, and which are similar to i. We have asked what things one is comparing when one is evaluating overall similarity, and have investigated the following alternatives:

(1) compare i and j in their temporal totalities,

(2) compare the history of j up to the time that A becomes true with the corresponding history of i,

(3) compare an A-slice of j with a current slice of i,

(4) compare an A-slice of j with a co-temporaneous slice of i, and

(5) compare a current slice of i with a co-temporaneous (current) slice of j.

Although each of these alternatives worked for some counterfactuals, none of them works for all of the counterfactuals which we have considered. Furthermore, there are some counterfactuals for which none of the alternatives works. The early warning device example and the example of Mort's weight are such cases. As is clear from the list above, we have not exhausted the possible alternatives. For example, we might try comparing the two closest slices of i and j, whenever they happen to take place, or (if this is different) the two closest co-temporaneous slices. We have selected for discussion only the most natural candidates, but none of the unnatural candidates which I have investigated does any better.

Could it be that while there is no single notion of similarity which gives us a uniform analysis of counterfactuals, that there is a range of notions of similarity such that (a) each counterfactual is correctly analysed by one notion of similarity in this range, and (b) the context of assertion indicates which in this range is the intended notion of similarity? In discussing Quine's examples,namely

> If Caesar had been in command [in Korea]he would have used the atom bomb
>
> versus
>
> If Caesar had been in command he would have used catapults,

Lewis says,

> In one context, we may attach great importance to similarities and differences in respect of Caesar's character and in respect of regularities concerning the knowledge of weapons common to commanders in Korea. In another context we may attach less importance to these similarities and differences, and more importance to similarities and differences in respect of Caesar's own knowledge of weapons. The first context resolves the vagueness of comparative similarity in such a way that some worlds with a modernized Caesar in command come out closer to our world than any with an unmodernized Caesar. It thereby makes the first counterfactual true. The second context resolves the vagueness in the opposite direction, making the second counterfactual true. Other contexts might resolve the vagueness in other ways. A third context, for instance, might produce a tie between the closest worlds with modernized Caesars and the closest worlds with unmodernized Caesars. That context makes both counterfactuals false.[12]

Could it be that our problem is similar, and that the context of assertion can be called on to indicate the appropriate notion of similarity?

Note first that if A is false, and if (A & B) and (A & $\overline{B}$) are both possible, by choosing appropriate similarity relations we can always make $A \mathrel{\Box\!\!\rightarrow} B$ true at i, or make $A \mathrel{\Box\!\!\rightarrow} \overline{B}$ true at i, or make neither true. To make $A \mathrel{\Box\!\!\rightarrow} B$ true, merely divide the set of worlds into two subsets, W_1 and W_2. Let W_1 be the set of worlds at which $A \rightarrow B$ is true, and let W_2 be the set of worlds at which $A \rightarrow B$ is false. Order the worlds so that every world in

W_1 is more similar to i than any world in W_2. Order worlds within W_1 and W_2 in any way you like. Then among A-worlds, any in which B is true will be more similar to i than any in which B is false, so $A \square\!\!\rightarrow B$ will be true at i. To make $A \square\!\!\rightarrow \overline{B}$ true at i, merely let W_1 be the set of worlds at which $A \rightarrow \overline{B}$ is true and W_2 the set of worlds at which $A \rightarrow \overline{B}$ is false. Finally, to make neither $A \square\!\!\rightarrow B$ nor $A \square\!\!\rightarrow \overline{B}$ true, let all worlds be equally similar to i. Thus, while it is true that for each counterfactual there is some measure of similarity which gives the correct truth value, this is trivially true; there is also a measure of similarity which gives us the incorrect truth value.

Secondly, the choice among the five (or more) alternative construals of similarity above does not seem to be a choice as to how to resolve vagueness. It does not seem to be a question, as in Quine's example, of how to weigh various respects of similarity. In that example, in the context of a discussion about military psychology and the relative tendencies of military leaders toward offense or defense, we might say that if Caesar were in Korea, he would have used the atom bomb. However, in discussing the ways that generals adapted their attack to the vagaries of terrain, contrasting Caesar's methods in Alesia with his methods in Britain, we might say that if Caesar were in Korea, he would have used catapults. Note that in the former case we would put emphasis on the word 'Caesar', whereas in the latter case we would place emphasis on the word 'Korea'. The problem we face here, however, is not a problem of context, for we may need to choose two different bases of comparison in the same context. Change the doomsday button example slightly, supposing the button to be an ancient, but still operational, artifact left by a lost culture of aliens. In showing the button to a group of journalists, a general might say, "If I pushed this button, the universe would be destroyed." As we have seen, to get the correct truth value for this counterfactual we see what happens in the world in which the general pushes the button whose history up to the time of the button push is most like our cotemporaneous history. But suppose the general now adds, "and if Attilla the Hun had pushed this button, the universe would have been destroyed." Now we go wrong if we compare histories up to the present; to get the correct truth value we must compare histories up to the time of Attilla. The choice of comparison basis has changed. But the context of utterance has remained the same. Hence it cannot be context which determines the choice of comparison basis.[13]

Thirdly, in some of our examples there is no plausible choice of comparison basis which seems to give the correct truth value. In the example of Mort's weight, any plausible choice of comparison basis forces either $A \Box\!\!\rightarrow B$ or $A \Box\!\!\rightarrow \bar{B}$ to be true, whereas we want both to be false. In the early warning device example, to make the counterfactual 'if Society Far were to receive a signal, the universe would be ⅔ destroyed' true in the given world *i*, we must find worlds in which a signal is received and compare them with *i* in respect to their histories up to the time the signal is *emitted*, not up to the time the signal is received. But change the example slightly (keep the context of utterance the same if you like) so that instead of emitting the signal 3 years before the cataclysm begins, the device emits a signal only after ½ of the universe has been destroyed, and in comparing histories up to the time the signal is emitted we will make the counterfactual false. To make it true we have to compare histories only up to some earlier time, some time before the cataclysm is sufficiently underway. For in the changed example, any world in which the signal is emitted and is genuine will be a world which is already ½ destroyed and which is therefore sufficiently different from the given world *i* that worlds in which the signal is emitted through malfunction will more similar. Note finally that there is no obvious relationship between the counterfactual 'if Society Far were to receive a signal, the universe would be ⅔ destroyed' (including its context of utterance) and the time we must stop comparing histories. In the unmodified example we stopped comparing histories at the time of emission of the signal; in the modified example we stopped comparing at some earlier time before the cataclysm was sufficiently underway. The choice of times does not seem to be determined by the counterfactual itself or by the context of utterance (the counterfactual and context were the same in both cases), but rather by the details of the working of the device. Thus there is no way in which the counterfactual along with its context of utterance determines an appropriate basis for comparison.

There is no question that, for a counterfactual hypothesis A, there are respects for which the propositions that would be the case if A were the case are just the propositions which are the case in those A-worlds most similar to ours in the given respects. $A \Box\!\!\rightarrow B$ will always be true just in case B is true in those A-worlds most similar to ours with respect to the *truths which would remain true if A were the case*. This much was clear from our discussion of the classical approach to counterfactuals. Unfortunately it is not generally the case that the A-worlds most like ours with respect to the truths which would remain truths if A were the case coincide with the A-worlds most like ours. In many cases the A-worlds which *would* be actual if A were true will be less similar to ours than other A-worlds which *could* be actual if A were true. When this is the case, the analysis of counterfactuals in terms of over-all similarity will give us the wrong truth-value. Thus in the EWD example, the world which would have been actual had a signal been received was less like the given world than other worlds in which the signal was received. Given the counterfactual hypothesis A, there are many ways to adjust the truth values of other propositions in order to accommodate A. In many cases, the adjustment which *would* be made if A were the case would bring about a world much less like ours than other worlds which could be brought about by less drastic adjustments. It is perhaps a brute fact that some changes would bring about more drastic over-all changes than they need bring about. There is certainly no reason to expect that the changes which *would* result from a change in the truth-value of A are the minimum changes compatible with logical necessity. But when the changes which would be brought about by A are more drastic overall then they need be, the similarity approach will give us the wrong answers.

Since there is some notion of similarity, namely similarity with respect to those truths which would remain truths if A were the case, which gives us the right answer for the counterfactual $A \Box\!\!\rightarrow B$, it is no surprise that the similarity approach seems to provide an acceptable logic for the counterfactual conditional. Unfortunately, there seems to be no non-circular way of explicating the required notion of similarity which does not itself involve counterfactual discourse. The problem is no different from that faced by the classical approach, wherein there seemed to be no non-circular way of explicating the required notion of co-tenability.

The similarity approach, understood as an analysis of the truth conditions or ordinary counterfactuals, requires that when we specify for the formal theory an interpretation whose set of possible worlds is the set of real possible worlds, and whose similarity relation is the real (but possibly vague) similarity relation on possible worlds, the resulting interpretation makes true just those counterfactuals which are, in fact, true.[14] Although for each counterfactual, there is trivially, a similarity relation which works for that counterfactual, there appears to be no similarity relation specifiable in non-counterfactual terms, which satisfies the constraint above. We conclude that the similarity approach fails to give an analysis of the truth conditions of ordinary counterfactuals.

References

[1] David Lewis, "Completeness and Decidability of Three Logics of Counterfactual Conditionals," *Theoria* 37(1971): 74-85.

[2] ______, *Counterfactuals*, (Harvard University Press, 1973).

[3] Donald Nute, "Counterfactuals," *Notre Dame Journal of Formal Logic* XVI, 4(October 1975): 476-482.

[4] ______, "Counterfactuals and the Similarity of Words [sic]," *Journal of Philosophy* LXXII, 21(1975): 773-8.

[5] Robert Stalnaker, "A Theory of Conditionals," in N. Rescher, *Studies in Logical Theory*, (Oxford: Blackwell, 1968).

[6] ______ and R. Thomason, "A Semantic Analysis of Conditional Logic," *Theoria* 36(1970): 23-42.

[7] Richmond Thomason, "A Fitch-Style Formulation of Conditional Logic," *Logique et Analyse* 52(1970): 397-412.

[8] Bas van Fraassen, "The Logic of Conditional Obligation," *Journal of Philosophical Logic* I(1972): 417-38.

Notes

[1] Earlier versions of this paper have been read to colloquia at the University of Michigan and the University of Massachusetts. I am grateful to the participants of those groups, and especially to John G. Bennett, for their many helpful suggestions.

[2] See David Lewis, [1, 2]; Robert Stalnaker, [5]; Robert Stalnaker and Richmond Thomason, [6]; Richmond Thomason, [7]; Bas van Fraassen, [8]. All subsequent references to Lewis will be to *Counterfactuals*, and all subsequent references to Stalnaker will be to "A Theory of Conditionals."

[3] Lewis observes the equivalence of the Classical theory with his under the assumption that the truths co-tenable with A are just those which (according to Lewis' theory) would remain truths if A were the case. Cf. Lewis, p. 69-70.

[4] Cf. Lewis, p. 48.

[5] Lewis, p. 75.

[6] Lewis, p. 75

[7] Of course the situation will vary depending on our individuation principle for temporal slices. Consider, for example, the fact that the paper which you are currently reading was written by me. Is that a fact about the current slice of the world? If so, then no world could have an identical current slice unless in that world, at some time in the past, I wrote this paper. If not, then a world in which somebody else wrote

this paper could have an identical current slice. I shall ignore such problems, although to the extent that we take temporal slices seriously, they are serious problems.

[8] This is not a possibility on Stalnaker's account.

[9] I shall ignore the considerable problems involved in cross-world identification of times.

[10] Remember that on Lewis' account $A \square\!\!\rightarrow B$ and $A \square\!\!\rightarrow \overline{B}$ are not contradictories.

[11] See Donald Nute, [3, 4]. This particular problem can be avoided within the framework proposed by Nute. In his formulation, $A \square\!\!\rightarrow B$ is true at i iff $A \rightarrow B$ is true at all worlds sufficiently similar to i to warrant consideration. In this example, it can be claimed that worlds in which Mort's weight is toward the top of the 300-400 range are also sufficiently similar to the given world to warrant consideration.

[12] Lewis, p. 67.

[13] Lewis uses an analogous argument to show that the counterfactual conditional cannot be a strict conditional whose strength depends on context of utterance. Cf. Lewis, p. 13.

[14] Naturally we must also specify values for the non-logical constants of the language.