Backward causation and the Stalnaker-Lewis approach to counterfactuals

Michael Tooley

Whether backward causation is logically possible is a deeply controversial matter, and one on which, in the present paper, I shall take no stand. The question to be considered is what relation, if any, there is between the logical possibility of backward causation and a Stalnaker-Lewis-style account of the truth conditions of counterfactuals, and the thesis that I shall be defending is that, if backward causation is logically possible, then a Stalnaker-Lewis-style account of the truth conditions of counterfactuals cannot be sound.

1. Counterfactuals and similarity relations over possible worlds

This general approach to counterfactuals – which appeals to similarity relations between possible worlds – was first set out by Robert Stalnaker (1968), and then a modified, and in some ways more satisfactory version, of it was advanced and defended by David Lewis (1973, 1979).

In his article, Stalnaker initially set out a very general account of counterfactuals in terms of the idea of a 'selection-function', an account that made no reference to similarity across possible worlds:

> … our semantical apparatus includes a *selection function, f*, which takes a proposition and a possible world as arguments and a possible world as its value. The *s*-function selects, for each antecedent *A*, a particular possible world in which *A* is true. The *assertion* which the conditional makes, then, is that the consequent is true in the world selected. A conditional is true in the actual world when its consequent is true in the selected world. (1968: 103)

What informal idea did Stalnaker want to capture by means of the concept of a selection function? Stalnaker's answer was that the informal truth conditions that he had proposed earlier in his article 'required that the world selected *differ minimally* from the actual world'. Thus, 'the selection is based on an ordering of possible worlds with respect to their resemblance to the base world' (1968: 104).

Analysis 62.3, July 2002, pp. 191–97.

Let us now turn to David Lewis's version of the general, similarity-across-possible-worlds approach to counterfactuals. As set out in Lewis (1973), his account was as follows:

> $\phi \,\Box\!\!\rightarrow \psi$ is true at a world i (according to a system of spheres \$) if and only if either
> (1) no ϕ-world belongs to any sphere S in $\$_i$, or
> (2) some sphere S in $\$_i$ does contain at least one ϕ-world, and $\phi \supset \psi$ holds at every world in S. (1973: 16)

But what is a 'system of spheres'? The basic idea here is that, for any possible world, all other possible worlds can be placed on spheres that are centred on the world in question, with the size of a given sphere representing how close each world on the sphere is to the world that lies at the centre of the given system of spheres. Thus, all worlds on a given sphere are equally similar to the world at the centre, and if one sphere is inside another, then the worlds on the inner sphere are more similar to the world at the centre than are the worlds on the outer sphere.

There is, accordingly, some ordering over worlds, and that relation is based upon some measure of overall similarity, and many readers of *Counterfactuals* interpreted Lewis as putting forward an account in which the overall similarity upon which the ordering was based was overall similarity *as judged by intuitive standards*. So interpreted, the approach is exposed to a decisive objection, advanced in early reviews by Jonathan Bennett (1974) and Kit Fine (1975), the thrust of which is that, so interpreted, the account would imply that counterfactuals such as the following get assigned the wrong truth-values:

(1) If Oswald had not killed Kennedy, someone else would have. (Bennett, 1974: 395)
(2) If Nixon had pressed the button, there would have been a nuclear holocaust. (Fine, 1975: 452)

Lewis, however, in his '*Postscripts to* "Counterfactual dependence and time's arrow" ', said that this reading was mistaken, and that his original account had not been formulated in terms of the idea that the relevant measure of similarity was that involved in our everyday, intuitive judgements of overall similarity. (1986: 52)

How, then, does one arrive at the relevant measure of similarity? Lewis's answer was as follows:

> The thing to do is not to start by deciding, once and for all, what we think about similarity of worlds, so that we can afterwards use those decisions to test Analysis 2. What that would test would be the combination of Analysis 2 with a foolish denial of the shiftiness of similarity. Rather, we must use what we know about the truth and falsity

> of counterfactuals to see if we can find some sort of similarity relation – not necessarily the first one that springs to mind – that combines with Analysis 2 to yield the proper truth conditions. It is this combination that can be tested against our knowledge of counterfactuals, not Analysis 2 by itself. In looking for a combination that will stand up to the test, we must use what we know about counterfactuals to find out about the appropriate similarity relation – not the other way around. (1979: 466–67)[1]

Lewis then attempted to show that one can provide an account of the factors that enter into relevant judgements of similarity between possible worlds in the case of counterfactuals that will generate the correct truth-values.

2. *A simple case of backward causation*

Let us now consider a world where backward time travel, and thus backward causation, is possible. Suppose, in particular, that there is a time machine, *M*, at a certain location *X* on Earth at time *t* in the year 2100, and which is in a state such that if its blastoff switch were to be flipped, then, provided that its fuel has not been removed, either at an earlier time, or at the very moment when the switch is flipped, it would travel forward in time to the year 2101, to location *Y* on Mars, where it would, at time *t**, both immediately remove all the fuel from any time machines in its neighbourhood, and flip their blastoff switches. Suppose, further, that there is another time machine, *N*, in the immediate vicinity of location *Y* at time *t** in the year 2101, and which is in a state such that if its blastoff switch were to be flipped, then, provided that its fuel has not been removed then or earlier, it would travel backward in time to the year 2100, to the immediate vicinity of location *X*, where it would, at time *t*, remove all the fuel from any time machines in its neighbourhood, and flip their blastoff switches. Consider, now, the following counterfactual:

(1) If the blastoff switch on time machine *M* were flipped at time *t*, it would wind up in location *Y* on Mars at time *t**, where it would remove fuel from all the time machines in its neighborhood, including time machine *N*, and flip their blastoff switches.

Our intuitive judgement here is surely that this counterfactual is true.

But, similarly, this is also the case with regard to the following counterfactual:

(2) If the switch on time machine *N* were flipped at time *t**, it would wind up in the immediate vicinity of location *X* on Earth at time

[1] The expression 'Analysis 2' refers to the type of analysis just set out.

> *t*, where it would remove fuel from all the time machines in its immediate neighborhood, including time machine *M*, and flip their blastoff switches.

The upshot is that, if backward time travel of the sort envisaged in this case is logically possible, then any similarity-across-possible-worlds account of counterfactuals cannot be correct unless the measure of similarity that is used makes it the case that the above two counterfactuals both turn out to be true.

3. Refining the case

The argument could now be formulated in terms of the case just mentioned, and, if one accepted the defence of time travel offered by David Lewis (1976), there could be no objection to using the above case. Moreover, the above case has the merit that it does not involve any causal loops, let alone causal loops of the two most problematic sorts – namely, that of the self-supporting variety, and, most dramatically, that of the self-undercutting variety.

Nevertheless, the world described above is one where there *could* be a self-undercutting causal chain. One need merely suppose that time machine *M* is programmed slightly differently, so that, on arriving at location *X* on Mars, it does not remove the fuel from any time machines in its vicinity. Then, when it flips the blastoff switch for time machine *N*, the latter will travel back to the vicinity of location *X*, at time *t*, where it will remove the fuel from time machine *M*, thereby preventing *M* from travelling forward to time t^*.

The problem, in short, is that one might hold that while there are some logically possible worlds that contain backward causation, there are no logically possible worlds where causal loops are nomologically possible. Let us shift, then, to a case where not only are there not, as a matter of fact, any causal loops, but where the laws of the world are such that causal loops cannot arise. Consider, in particular, a world where the basic individuals are locations and moments of time, and where there are only two properties – *P* and *Q* – that a location can have at a given time, and only two laws, one a forward causal law, and the other a backward causal law:

> Law 1: For any location x, and time t, if location x has both property *P* and property *Q* at time t, then that state of affairs causes a related location $x + \Delta x$ to have property *P*, and to lack property *Q*, at the later time $t + \Delta t$.
>
> Law 2: For any location x, and time t, if location x has both property *P* and property *Q*, at time t, then that state of affairs causes a related location $x - \Delta x$ to have property *P*, and to lack property *Q*, at the earlier time $t - \Delta t$.

Suppose, finally, that world in question is as follows:

	World W_0	
Times:	t	$t + \Delta t$
States of Affairs:	x lacks property P	$(x + \Delta x)$ lacks property P
	x has property Q	$(x + \Delta x)$ has property Q

Consider, then, first of all, how the world would have been if location x had had property P at time t. The correct answer, surely, is that if x had had property P at time t, then it would have had both property P and property Q at time t, and so, in view of Law 1, location $x + \Delta x$ would have had property P, but would not have had property Q, at time $t + \Delta t$. For, other than the fact that laws have been introduced to make causal loops nomologically impossible, the present case parallels exactly the time travel case described in the previous section: property P corresponds to the property of having a blastoff switch that has been flipped, while property Q corresponds to having fuel tanks that have not been emptied.

The world that would have existed if location x had had property P at time t is, accordingly, as follows:

	World W_1	
Times:	t	$t + \Delta t$
States of Affairs:	x has property P	$(x + \Delta x)$ has property P
	x has property Q	$(x + \Delta x)$ lacks property Q

Next, consider how the world would have been if location $x + \Delta x$ had had property P at time $t + \Delta t$. The correct answer, surely, is that if $x + \Delta x$ had had property P at time $t + \Delta t$, then it would have had both property P and property Q at time $t + \Delta t$, and so, in view of Law 2, location x would have had property P, but would not have had property Q, at time t. So the world that would have existed if location $x + \Delta x$ had had property P at time $t + \Delta t$ is as follows:

	World W_2	
Times:	t	$t + \Delta t$
States of Affairs:	x has property P	$(x + \Delta x)$ has property P
	x lacks property Q	$(x + \Delta x)$ has property Q

The argument can now be put as follows. First, both of the following counterfactuals are true in world W_0:

(1*) If location x had had property P at time t, then location $x + \Delta x$ would not have had property Q at time $t + \Delta t$;
(2*) If location $x + \Delta x$ had had property P at time $t + \Delta t$, then location x would not have had property Q at time t.

Secondly, let us say that a world is an A-world if and only if it is a world where x has property P at time t, and a B-world if and only if $x + \Delta x$ has property P at time $t + \Delta t$. If a Stalnaker-Lewis-style account of the truth conditions of counterfactuals is correct, it follows that counterfactual (1*)

cannot be true unless either world W_2 is not an A-world, or else world W_1 is closer to world W_0 than W_2 is. But W_2 is an A-world. Therefore, if a Stalnaker-Lewis-style account of the truth conditions of counterfactuals is correct, it follows that counterfactual (1*) cannot be true unless world W_1 is closer to world W_0 than W_2 is.

Thirdly, and similarly, if a Stalnaker-Lewis-style account of the truth conditions of counterfactuals is correct, counterfactual (2*) cannot be true unless either world W_1 is not a B-world, or else world W_2 is closer to world W_0 than W_1 is. But W_1 is a B-world. Therefore, if a Stalnaker-Lewis-style account of the truth conditions of counterfactuals is correct, it follows that counterfactual (2*) cannot be true unless world W_2 is closer to world W_0 than W_1 is.

Thus, if a Stalnaker-Lewis-style account of the truth conditions of counterfactuals is correct, it follows that counterfactuals (1*) and (2*) cannot both be true unless it is true both that world W_1 is closer to world W_0 than W_2 is, and that world W_2 is closer to world W_0 than W_1 is. This, however, is impossible.

Accordingly, if backward causation, at least of the non-looping variety, is logically possible, then either a Stalnaker-Lewis-style account of the truth conditions of counterfactuals is unsound, or else counterfactuals (1*) and (2*) cannot both be true. But if backward causation of the type in question is logically possible, then both (1*) and (2*) are true. Therefore, if backward causation, at least of the non-looping variety, is logically possible, a Stalnaker-Lewis-style account of the truth conditions of counterfactuals cannot be correct.

4. *Summing up*

Can one define a relation of similarity across possible worlds that can be used in a Stalnaker-Lewis-style account of the truth conditions of counterfactuals, and that will generate the correct truth-values for all counterfactuals? What I have argued here is that if backward causation – at least of the non-looping variety – is logically possible, then this cannot be done. For one can then find a logically possible world – containing backward causation – and a pair of counterfactuals that are true in that world, but whose truth would, given a Stalnaker-Lewis-style account of counterfactuals, entail logically incompatible propositions concerning the relative closeness of two other possible worlds to the world in question. So if backward causation is logically possible, counterfactuals cannot be analysed in a Stalnaker-Lewis fashion.

The University of Colorado at Boulder
Boulder, CO 80309, USA
Michael.Tooley@Colorado.edu

References

Bennett, J. 1974. Counterfactuals and possible worlds. *Canadian Journal of Philosophy* 4: 381–402.

Fine, K. 1975. Critical notice of Lewis 1973. *Mind* 84: 451–58.

Lewis, D. 1973. *Counterfactuals*. Cambridge, Massachusetts: Harvard University Press.

Lewis, D. 1976. The paradoxes of time travel. *American Philosophical Quarterly* 13: 142–52. Repr. in his *Philosophical Papers*, II, 67–80. New York: Oxford University Press, 1986.

Lewis, D. 1979. Counterfactual dependence and time's arrow. *Noûs* 13: 455–76. Repr. in his *Philosophical Papers*, II, 32–52. New York: Oxford University Press, 1986.

Lewis, D. 1986. *Postscripts to* 'Counterfactual dependence and time's arrow'. In his *Philosophical Papers*, II, 52–66. New York: Oxford University Press.

Stalnaker, R. C. 1968. A theory of conditionals. In *Studies in Logical Theory*, ed. N. Rescher, 98–112. Oxford: Blackwell. Repr. in *Causation and Conditionals*, ed. E. Sosa, 165–79. Oxford: Oxford University Press, 1975.