Tooley on backward causation

Paul Noordhof

Michael Tooley has argued that, if backward causation (of a certain kind) is possible, then a Stalnaker-Lewis account of the truth conditions of counterfactuals cannot be sound. I shall argue that he is wrong.[1] According to David Lewis,

> A counterfactual 'If it were that A, then it would be that C' is non-vacuously true if and only if some (accessible) world where both A and C are true is more similar to our actual world, over-all, than is any world where A is true but C is false. (Lewis 1979: 41)

Lewis's criteria for assessing the similarity between possible worlds are as follows.

> (A) It is of the first importance to avoid big, widespread, diverse violations of law.
> (B) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
> (C) It is of the third importance to avoid even small, localized, simple violations of law.
> (D) It is of little or no importance to secure approximate similarity of particular fact, even in matters which concern us greatly. (Lewis 1979: 47–48)

The basic idea is that we are to consider all those worlds in which A is true. The counterfactual will be true if the worlds in which C is also true are more similar according to the criteria laid out in (A) to (D) than any world in which C is not true. The crucial point is that which close worlds we consider is fixed by, in the first instance, the envisaged truth of the antecedent.

Tooley invites us to imagine a world in which the following holds.

> 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 $x + \triangle x$ to have property P, and to lack property Q, at the later time $t + \triangle t$.

[1] Initially, Tooley suggests that he shall demonstrate that backward causation is incompatible with the Stalnaker-Lewis style account of the truth conditions of counterfactuals (Tooley 2002: 191). By the end, it becomes clear that he only thinks that the Stalnaker-Lewis style account of the truth conditions of counterfactuals is incompatible with backward causation in worlds whose laws rule out causal loops.

Analysis 63.2, April 2003, pp. 157–62. © Paul Noordhof

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 $x - \triangle x$ to have property P, and to lack property Q, at an earlier time $t - \triangle t$.

According to the story, the world in which we are located, w_0, has the following characteristics.

World W_0

Times	t	$t + \triangle t$.
States of affairs	Not-Px, Qx	Not-P$(x + \triangle x)$, Q$(x + \triangle x)$

Consider the counterfactuals

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

In order for the first counterfactual to be true, the following should hold in the closest worlds *in which location x has property P at time t.*

World W_1

Times	t	$t + \triangle t$
States of affairs	Px, Qx	P$(x + \triangle x)$, not-Q$(x + \triangle x)$

Given the laws which hold, the counterfactual is plausible. By Law 1, if at t, Px and Qx, then at $t + \triangle t$, P$(x + \triangle x)$ and not-Q$(x + \triangle x)$. By the same token, for the second counterfactual to be true, the following should hold in the closest worlds *in which location $x + \triangle x$ has property P at time $t + \triangle t$.*

World W_2

Times	t	$t + \triangle t$
States of affairs	Px, not-Qx	P$(x + \triangle x)$, Q$(x + \triangle x)$

This also seems plausible, by Law 2. If, at $t + \triangle t$, P$(x + \triangle x)$, Q$(x + \triangle x)$ then at some time $\triangle t$ earlier, at a place $\triangle x$ distance from x, one would expect Px and not-Qx.

Tooley claims that '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 W_0 than W_2 is, and that W_2 is closer to W_0 than W_1 is. This, however, is impossible' (Tooley 2002: 196). Hence the approach cannot capture what we want to say about this case.

Tooley seems to ignore the fact that the two counterfactuals have different antecedents concerning the instantiation of P at different times and places. These differences imply that the truth of each antecedent would select (at the outset) different spheres of possible worlds with a potentially different order of similarity. In order to assess (1*) we would be looking at the Pxt-worlds and asking whether worlds in which not-Q($x + \triangle x$) at $t + \triangle t$ are closer than worlds in which Q($x + \triangle x$) at $t + \triangle t$. In order to assess (2*), we would be looking at P($x + \triangle x$)-worlds and asking whether worlds in which not-Qx at t are closer than worlds in which Qx at t. Now it is, of course, true that W_1 and W_2 are members of both sets of worlds as specified (I shall come back to this point in a moment). Nevertheless, it does not follow from this that they will be picked out as the closest in both cases. Instead, W_1 may be closer amongst the Pxt-worlds and W_2 may be closer amongst the P($x + \triangle x$)-worlds. In order to establish his conclusion, Tooley needed to establish that this could not reasonably be claimed. But all he seems to do is take it that the truth of (1*) and (2*) requires that both W_1 is closer than W_2 and W_2 closer than W_1 *tout court*.

The point I am making can be illustrated by a case with none of the peculiar features of Tooley's. Suppose, in fact, I did not sell my gun last week, stalked a man and shot him. Consider the following fore-tracking and back-tracking counterfactuals:

(3) If I had sold the gun last week, I would not have shot him.
(4) If I had not shot him, I would have sold the gun last week.
(5) If I had sold the gun last week and agreed to kill him, I would not have shot him.

Let W_3 be a world in which I sold the gun and I did not shoot him. For (3) to be true, W_3 (or worlds like it) are the closest Gun-Selling-worlds. It is also the case that W_3 is a member of both the Gun-Selling-worlds and the Didn't-Shoot-worlds. However, I take it, W_3 (or worlds like it) need not be the closest Didn't-Shoot worlds. By the time of the shooting, it is already established that I didn't sell the gun. So, to maximize perfect match the closest worlds in which I did not shoot him would be ones in which I would not have sold the gun. I may just have decided better of it (say). That's why we don't, in general, think that (4) is true. Similarly, though W_3 is a member of both the Gun-Selling-worlds and the Gun-Selling-&-Agreed-to-Kill-worlds, I take it we would not conclude that (5) must be false. Indeed, it seems true. It would be entirely inappropriate to argue that those committed to the truth of (3) and the falsity of (4) are committed to the impossibility that W_3 (or worlds like it) are both the closest worlds and not the closest worlds. Nor are matters different regarding whether one world is closer to another. Let W_4 be a world in which I sell the gun, buy it back and decide to shoot him. I take it that W_3 is closer than W_4 in the set of Gun-Selling-Worlds. However, I take it that W_4 is closer than W_3 in the Gun-Selling-&-Agreed-to-Kill-worlds.

The proponents of the Lewis-Stalnaker approach to counterfactuals would, thus, treat Tooley's case in the same way as others. They would note with interest our intuitions with regard to (1*) and (2*) and argue that this showed something either about the assumed context at work for each or that Lewis's similarity weighting needs adjustment.

According to the first option, when we are presented with the case, we assume that Qx at t for (1*) and $Q(x + \triangle x)$ at $t + \triangle t$ for (2*). Since we have made these assumptions, W_2 does not even fall into the Px-worlds (really Px & Qx worlds) and W_1 does not even fall into the $P(x + \triangle x)$-worlds (really $P(x + \triangle x)$ & $Q(x + \triangle x)$ worlds). That's why there is no problem. The verdicts at which we arrive concerning (1*) and (2*) strike us as plausible because the way in which the case has been described encourages us to fill out the antecedent in the way indicated.

Suppose we try to envisage a situation in which this context is not at work for (1*). What would Lewis's similarity weighting proclaim? We are to suppose that location x has property P at time t. Given that our world is W_0 and, assuming determinism, one way in which this could happen is by a small miracle occurring just prior to t. We would thereby maximize perfect match up to a moment before t. We then roll the world on, according to the laws, in particular, Law 1, to $t + \triangle t$, and conclude that $x + \triangle x$ would lack Q. Alternatively, we could have a small miracle just before $t + \triangle t$, to instantiate $P(x + \triangle x)$. Then, by Law 2, we would have Px (and not Qx) at t. Law 1 would not apply. I take it that these two options are on a par regarding law violations (or, at least, this can't be ruled out). The second option allows perfect match right up to t. The first loses it just before t. So, by Lewis's similarity weighting, (1*) is false.

Let's turn to (2*). Again, there are two options. According to the first, we might imagine a small miracle occurring just before $t + \triangle t$ so that $P(x + \triangle x)$ holds. In such circumstances, we would then roll the world back according to the laws, in particular, Law 2, to t and conclude that Px and not-Qx. Of course, an alternative would be to suppose that, by some miracle just before t, Px and Qx at t. Then we would get, via Law 1, P and not-Q at $x + \triangle x$. Law 2 would not apply. I take it once more that these two options are on a par regarding law violations. The first option allows for perfect match right up to t. The second option loses it just before t. According to Lewis's similarity weighting (2*) is true.

Do our intuitive assessments of these counterfactuals free of the assumed context I mentioned agree with this asymmetry? In so far as I can get this issue into clear focus, my inclination is to say ‘no’. This suggests that we may have learnt something important about the formulation of the perfect match clause and also about the final clause of Lewis’s similarity weighing. Normally, ensuring perfect match close up to the time of the antecedent is a way of ensuring that the circumstances in which we envisage the antecedent to hold will be very similar to the actual circumstances. However, in Tooley’s case involving backward causation, this is not so. If you maximize perfect match, you get rid of Q*xt*. I suggest the following adjustment.

> (B*) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails unless one reduces approximate match around the time of the circumstances mentioned in the antecedent.[2]

This alone would still entail that (1*) was false. It wouldn’t be the case that *location x* + △*x would not have had property Q at time t* + △*t* because it *might* be the case that not-Q*x* at *t*. So we should add

> (D*) It is of fourth importance to maximize similarity of independent particular fact closest to the time of the circumstances mentioned in the antecedent.

The proper characterization of ‘independent particular fact’ is a matter of some delicacy which cannot be given in the space available. Let me just note that, in the present case, two facts are independent if they don’t stand in the ancestral of *counterfactual dependence to each other where *counterfactual dependence is characterized by possible worlds with a similarity weighing given by (A) to (C) (but not (D*)). Thus Q*x* at *t* is independent of P*x* at *t*, because if P*x* at *t* were the case, it might still be that not-P(*x* + △*x*) at *t* + △*t*.

The adjustments just indicated have an interesting consequence. The closeness of possible worlds is no longer an absolute matter with the antecedent merely selecting the appropriate subset that needs to be considered. Instead, closeness of possible worlds becomes relative to the antecedent in question. This is a significant point of difference between the Gun-shooting cases and the cases for which we have reason to thank Michael Tooley.[3]

[2] Actually, I believe this is in need of further adjustment to deal with indeterministic cases, but I bracket such matters here.

[3] My thanks to Helen Beebee and Michael Clark, who made this paper very much better than it would otherwise have been.

University of Nottingham,
University Park, Nottingham NG7 2RD, UK
Paul.Noordhof@nottingham.ac.uk

References

Lewis, D. 1979. Counterfactual dependence and time's arrow. *Noûs* 13: 455–76. Repr. in his *Philosophical Papers*, Vol. 2, 32–66. Oxford: Oxford University Press, 1986. [Page references in the text are to the latter.]

Tooley, M. 2002. Backward causation and the Stalnaker-Lewis approach to counterfactuals. *Analysis* 62: 191–97.