Counterfactuals, causal independence and conceptual circularity

Jonathan Schaffer

David Lewis's semantics for counterfactuals remains the standard view. Yet counter-examples have emerged, which suggest a need to invoke causal independence, and thus threaten conceptual circularity. I will review some of these counter-examples (§§1–2), illustrate how causal independence proves useful (§3), and suggest that any resulting circularity is unproblematic (§4).

Analysis 64.4, October 2004, pp. 299–309. © Jonathan Schaffer

1. Lewis on counterfactuals

Lewis (1973a, 1973b, 1979, 1986b) develops what remains the standard semantics for counterfactual conditionals. Suppose there are possible worlds,[1] ordered by comparative similarity.[2] Then Lewis's semantics assigns truth-conditions as follows:

(L) $p \supset q$ is true at w iff: if there are p-worlds, then there is a $p\&q$-world closer to w then any $p\&\sim q$-world.

L proves to fit intuitions remarkably well. Though, of course, there are counter-examples.

One counter-example is Kit Fine's (1975) case of Nixon and the bomb:

(Fine's bomb) At w_0, Nixon does not press the nuclear button, so no nuclear holocaust occurs. The following counterfactual seems intuitively true at w_0: 'If Nixon has pressed the button, then there would have been a nuclear holocaust.'

Intuitively, a holocaust would make for a vast dissimilarity. Which would seem to force L to rule the relevant counterfactual false at w_0: there are (pressing&holocaust)-worlds, but there are (pressing&holocaust-averted)-worlds that are closer.

Lewis (1979) replies to Fine's bomb by explicating a system of similarity weights, under which L rules the relevant counterfactual as true:

(1) It is of the first importance to avoid big miracles.
(2) It is of the second importance to maximize the region of perfect match.
(3) It is of the third importance to avoid small miracles.
(4) It is of the fourth importance to maximize the region of imperfect match.

Lewis then compares a (pressing&holocaust)-world, w_1, to (pressing&no-holocaust)-worlds including w_2 where the holocaust is averted but traces of Nixon's pressing propagate, and w_3 where all traces of Nixon's pressing are deleted. Lewis needs w_1 to come out closer to w_0 than w_2 or w_3 does.[3] The weighting system of 1–4 delivers this. First, w_1 comes out closer to w_0 than w_2 does, by the priority of 3 over 4: while w_2 buys imperfect

[1] Never mind whether the worlds are ontologically primitive, reducible, fictional or whatnot. It only matters that quantification over possible worlds is permitted.

[2] This weak ordering can be pictured as a system of spheres centered on actuality. See Lewis 1973a: 14 for an elegant presentation of the formal requirements here.

[3] More generally, Lewis needs w_1 to come out closer to w_0 than any (pressing&no-holocaust)-world does. Here I am just laying the groundwork.

match with respect to the holocaust-free future, it costs at least a small miracle for the signal to the bomb to fizzle. Second, w_1 comes out closer to w_0 than w_3 does, by the priority of 1 over 2: while w_3 buys perfect match with respect to the holocaust-and-trace-free future, it costs a big miracle to erase all the traces of Nixon's action.

A second counter-example is the indeterministic version of Fine's bomb:

> (**Indeterministic bomb**) At indeterministic w_4, Nixon does not press the nuclear button, so no nuclear holocaust occurs. The following counterfactual seems intuitively true at w_4: 'If Nixon has pressed the button, then there might have been a nuclear holocaust.'

The problem is that there is an indeterministic (pressing&no-holocaust)-world w_5, where the holocaust is averted due to the chancy signal fizzling. And by 1–4, w_5 comes out closer to w_4 than any (pressing&holocaust)-world does, since w_5 costs no miracles whatsoever, and still buys the added match of a holocaust-free future.

Lewis (1986b) replies to Indeterministic bomb by extending the notion of a law violation to include a 'quasi-miracle', where chance outcomes seem 'to conspire to produce a pattern' (1986b: 60). The system of weights may then be amended as follows:

(1′) It is of the first importance to avoid big miracles or big quasi-miracles.
(2′) It is of the second importance to maximize the region of perfect match.
(3′) It is of the third importance to avoid small miracles or small quasi-miracles.
(4′) It is of the fourth importance to maximize the region of imperfect match.

Now similar reasoning to that employed in defusing Fine's bomb will defuse Indeterministic bomb. To buy perfect match with respect to the holocaust-free future would cost a big quasi-miracle; to buy imperfect match would cost a small quasi-miracle.

Never mind what exactly 'big' and 'small' miracles are, what exactly a 'quasi-miracle' is, what exactly 'maximization' of region of match consists in, or whether further worlds should be considered. So far I am merely bringing the system of weights on stage.

2. *Further counter-examples*

As ingenious as L plus 1′–4′ may be, further counter-examples lurk. One counter-example, due to Ryan Wasserman (manuscript), modifies Fine's bomb to circumvent Lewis's weighting system:

> (**Wasserman's Beetle**) w_6 is much like w_0, plus a sealed box containing a beetle, a button, and a wire leading out to the bomb. The wire is the only causal path out of the box. The beetle never bumps the button, so no holocaust occurs. Shortly thereafter, the entire box is cleanly erased. The following counterfactual seems intuitively true at w_6: 'If the beetle had bumped the button, then there would have been a nuclear holocaust.'[4]

Recall that in Fine's bomb, Lewis ruled the (pressing&perfect-convergence)-world w_3 more distant than the (pressing&holocaust)-world w_1, on the ground that the traces of the pressing would be widespread and diverse, costing a big miracle to erase. Wasserman's beetle is designed so that the traces are sealed in the box, and cleanly erased shortly thereafter. Perfect reconvergence no longer costs a big miracle. Thus, L plus 1′–4′ cannot respect the relevant counterfactual. There is a (pressing& perfect-reconvergence)-world w_7 that buys perfect match with respect to the holocaust-free future, while only costing a small miracle.[5]

A second counter-example, devised by Adam Elga (2000), shows that Lewis's weighting system does not handle anti-entropic processes, which achieve perfect convergence through a small miracle:

> (**Elga's anti-entropic processes**) At w_8, Greta cracks an egg on the pan, and soon has a fried egg. The following counterfactual seems intuitively true at w_8: 'If Greta had not cracked an egg on the pan, then she would not have had a fried egg.'

Lewis needs the (no-crack&no-fried-egg)-world w_9 to be closer to w_8 than any (no-crack&fried-egg)-worlds. But there is a world w_{10}, where Greta does not crack an egg, but a fried egg forms by an anti-entropic process (slime unrots into an egg and then warms – the temporal inverse of a fried egg cooling and then rotting). What Elga shows is that the smallest of miracles suffices to get w_{10} to converge on w_8 with respect to the future.[6] It would seem that w_9 and w_{10} are equally close to w_8. Each costs a small

[4] See Michael Tooley 2003 for a similar style of counter-example.

[5] Lewis would likely have dismissed Wasserman's beetle as *too farfetched*. Lewis says that ordinary thought presupposes the ubiquity of traces (1986b: 66), and that examples that conflict with ordinary presuppositions are generally unreliable: 'spoils to the victor!' (1986c: 203). But I find it doubtful that ordinary thought makes such a sophisticated presupposition about traces. Moreover, I see no reason why one cannot sometimes peer beyond ordinary presuppositions, provided one is careful. In any case, I think I grasp Wasserman's beetle well enough, and have clear enough intuitions.

[6] This is a consequence of the way anti-entropic processes are distributed in phase space.

miracle (in w_9, the shift to no-crack; in w_{10}, the shift in phase space to an entropic point); and each buys some perfect match (in w_9, perfect match in the past prior to the small miracle; in w_{10}, perfect match in the future after the small miracle; suppose these are equal in extent). Thus L plus 1′–4′ cannot respect the relevant counterfactual.[7]

A third counter-example, due to Sidney Morgenbesser (Michael Slote 1978), involves a bet on an indeterministic coin flip:

> (**Morgenbesser's coin**) At indeterministic w_{11}, while the coin is in mid-air, Lucky bets heads. The coin lands tails, so Lucky loses. The following counterfactual seems intuitively true at w_{11}: 'If Lucky had bet tails, he would have won.'[8]

Now L and 1′–4′ entail that the relevant counterfactual is false: Lucky merely *might* have won. To see this, compare the (Lucky-bets-tails&coin-lands-tails)-world w_{12}, with the (Lucky-bets-tails&coin-lands-heads)-world w_{13}. Given 1′–4′, w_{12} and w_{13} will come out equidistant from w_{11}. Each costs the same miracle of Lucky betting tails. Each costs perfect match with actuality from then on, neither costs any further miracles, and each buys an aspect of imperfect match – w_{12} preserves the outcome of the flip (tails), while w_{13} preserves the outcome of the bet (unlucky). Either might have the wider ramifications – for instance, either might inspire Nixon to press the button.[9] Since w_{12} and w_{13} come out equidistant, the relevant counterfactual is false on L.[10]

A second counter-example, developed by John Hawthorne (forthcoming; further discussed in Wasserman (manuscript)), explores the downside of Lewis's invocation of quasi-miracles against Indeterministic bomb:

> (**Hawthorne's compulsive**) At indeterministic w_{14}, Fred is a compulsive coin flipper, about to stage a million flips. But Fred is struck by

[7] Lewis acknowledged Elga's counter-example and considered revisions, but to my knowledge he did not decide on a remedy.

[8] See Igal Kvart 1986 and Stephen Barker 1999 for further exploration of this and related sorts of cases.

[9] Lewis 1979 mentions Morgenbesser's coin as a case in which imperfect match may help. The main text shows, though, that the story can be rigged so that imperfect match favours either w_{12} or w_{13}. Or neither.

[10] Lewis might bite the bullet here, as Slote did, and insist that the relevant 'would' counterfactual is false. He might dismiss our intuitions as vestiges of a deterministic mindset. I would agree with Dorothy Edgington that this would be 'wishful thinking' (2004: 17). Intuitions here may be buttressed by contrasting the supposition of Lucky switching his bet, with the supposition of Lucky swatting the coin in mid-air. On the latter supposition, we *do* intuit that Lucky merely *might* have won. There is a strong intuitive difference here: one suggestive of the role of causal independence in our assessment of counterfactuals.

lightning and dies. The following counterfactual seems intuitively true at w_{14}: 'If Fred had not been struck by lightning, he might have flipped all tails.'

Now L and 1′–4′ entail that the relevant counterfactual is false. There is a world w_{15} in which Fred flips all tails. But there is also a world w_{16} in which Fred produces some unremarkable outcome sequence σ (for instance, let σ begin: HHTHTTTTH ...) Now w_{16} will count as closer to w_{14} than w_{15} does, since w_{15} costs the quasi-miracle of all tails.

Things get worse. There is a non-zero chance ($1/2^{1000000}$) that Fred would have flipped all tails. This means that Lewis must assign *non-zero chances* to quasi-miracles, while ruling that they *would not* occur. This violates a plausible chance-entails-might principle:

(CEM) If $ch_{w,t}(p) > 0$, then $might_{w,t}$: p

Things get worse still. There is a greater chance (by 2 : 1) that Fred would have flipped either all heads or all tails, than that he would have produced σ. But Lewis is committed to saying that Fred might have produced σ, but would not have produced either all heads or all tails. This violates a plausible principle that one might think of as a *chance-might penumbral connection* principle:

(CMPC) If $ch_{w,t}(p) > ch_{w,t}(q)$, then (if $might_{w,t}$: q, then $might_{w,t}$: p)[11]

I will add one further counter-example, which combines Morgenbesser's coin with Hawthorne's compulsive:

(**Combination**) At indeterministic w_{17}, Fred is a compulsive coin flipper, and Lucky a compulsive gambler. Lucky bets that Fred will not flip a million tails in a row. Fred flips a million tails in a row, so Lucky loses the bet. The following counterfactual seems intuitively true at w_{17}: 'If Lucky had bet all tails, he would have won.'

But given L and 1′–4′, had Lucky bet all tails, what follows is that he *would not* have won, since the quasi-miracle of all tails would not have occurred. (One doesn't even get that Lucky *might* have won.) More carefully, compare w_{18} where Lucky bets all tails and Fred flips all tails,

[11] Lewis has a reply, based on distinguishing a *not-would-not* sense from a *would-be-possible* sense of 'might': $might_{nwn} =_{df} \sim(p > \sim q)$; $might_{wbp} =_{df} p \supset \Diamond q$ (1986a: 61–65). Lewis can then still affirm that Fred $might_{wbp}$ flip all tails. (Since $might_{wbp}$ embeds a diamond, all it requires is that at the 'nearest possible' p-world, the quasi-miracle q counts as *possible*.) But I see no linguistic evidence for this alleged ambiguity. In any case, Lewis would still be committed to a true interpretation of: 'If Fred had not been struck by lightning, he would not have flipped all tails.' Postulating an ambiguity in 'might' *cannot undo this*.

with w_{19} where Lucky bets all tails and Fred produces random sequence σ. Both w_{18} and w_{19} cost the same miracle of Lucky switching his bet to all tails. But w_{18} costs a further big quasi-miracle. So w_{19} comes out closer than w_{18}, and so by L, if Lucky had bet all tails, he would not have won. Which seems backwards.

3. *Causal independence*

These counter-examples all suggest a need to invoke *causal independence*.[12] Here is one way to express this idea: only match among *those facts causally independent of the antecedent* should count towards similarity. Not all matching is equal. After all, if outcome *o* causally depends on *p* or ~*p*, then *o* should be expected to *vary* with *p* or ~*p* – its varying should hardly count for *dis*similarity.

One can implement the idea of causal independence within L, by retreating back to 1–4, and amending 2 and 4 with a causal independence proviso:

(1c) It is of the first importance to avoid big miracles.
(2c) It is of the second importance to maximize the region of perfect match, *from those regions causally independent of whether or not the antecedent obtains*.
(3c) It is of the third importance to avoid small miracles.
(4c) It is of the fourth importance to maximize the spatiotemporal region of approximate match, *from those regions causally independent of whether or not the antecedent obtains*.

One thing nice (and perhaps novel) about this way of invoking causal independence is that it preserves L. It is conservative with respect to the standard semantics.[13,14]

Combining L with 1c–4c resolves the seven counter-examples above. These prove to turn on match from causally dependent regions, which 1c–4c is rigged to ignore.

[12] Kvart 1986, Barker 1999, and Edgington 2004 are among those who have offered a similar diagnosis. Thus Edgington speaks of 'the crucial role of causal independence' (2004: 18), adding (in reference to Morgenbesser's-coin style cases): 'I don't see how Lewis can handle these examples without appealing to the notion of causal independence' (2004: 20).

[13] Other causal independence proposals, such as Kvart 1986 and Barker 1999, require radical departure. These departures also tend to lose the generality of L, especially with counter-legals.

[14] As Adam Elga pointed out to me, 1c–4c involve a shift from *absolute closeness*, to *antecedent-relative closeness*. This matters when it comes to the validity of principles such as Substitution: $(A > B), (B > A), (A > C) \vdash (B > C)$. See Jonathan Bennett 2001 (194–96, 198–201) for further discussion, and a defence of antecedent relativity.

Starting with Fine's bomb, the (pressing&reconvergence)-worlds w_2 and w_3 were worrisome, because they bought match (imperfect and perfect, respectively) with respect to the holocaust-free future. But the holocaust-free future comprises a region of match that *causally depends* on whether or not the button is pressed. So on 1c–4c such match counts for nothing, and so cannot counterbalance the added convergence miracles of w_2 and w_3. Thus, w_1 is closer to w_0 than is w_2 or w_3. Which solves the problem.

Turning to Wasserman's beetle, note that the solution to Fine's bomb requires no assumption about the ubiquity of traces. So that solution will still work here. The (pressing&perfect-reconvergence)-world w_7 still buys perfect match with respect to the holocaust-free future, while costing a small miracle. It is just that what is bought now counts for nothing. Only the cost remains.

Moving to Elga's anti-entropic processes, recall that the anti-entropic world w_{10} was worrisome, because it bought perfect match with respect to the future. But the future converged on causally depends on whether or not Greta cracks an egg. So on 1c–4c, both w_{10} and the (no-crack&no-fried-egg)-world w_9 cost a small miracle, and each buys some perfect match: perfect match in the past prior to the small miracle in w_9; perfect match in the future after the small miracle in w_{10}. It is just that only w_9 buys match that counts.

Continuing over to Indeterministic bomb, the problem there concerned w_5, where the holocaust is averted by the chancy signal fizzling. This buys some match over any (pressing&holocaust)-world, at no cost in miracles. But once again this match is from a region causally dependent on whether or not the antecedent obtains. So on 1c–4c, both w_5 and the (pressing&holocaust)-world come out equally close to w_4. Which entails that either *might* have resulted. Which is what was wanted.

Shifting to Hawthorne's compulsive, note that the solution to Indeterministic bomb requires no distancing by, much less mention of, quasi-miracles (this is why I formulated 1c–4c by first retreating to 1–4). So the (all-tails)-world w_{15} and the (random-sequence)-world w_{16} will come out equally close to w_{14}, each costing the same miracle of Fred dodging lightning, and neither buying more match. Which entails that either *might* have resulted. Which is what was wanted, and which fits CEM and CMPC.

Flipping to Morgenbesser's coin, recall that the problem was that the (Lucky-bets-tails&coin-lands-tails)-world, w_{12}, and the (Lucky-bets-tails&coin-lands-heads)-world, w_{13}, came out equidistant from w_{11}. Each buys a different aspect of imperfect match – w_{12} preserves the outcome of the flip (tails), while w_{13} preserves the outcome of the bet (unlucky). But there is a causal difference – the outcome of the flip is causally

independent of Lucky's bet,[15] while the outcome of the bet is not. Hence w_{12} now comes out closer to w_{11} than w_{13} does. And so had Lucky bet tails, he would have won.

Finishing with Combination, the solutions to Hawthorne's compulsive and Morgenbesser's coin apply. As with Morgenbesser's coin, given that the outcome of the million flips are causally independent of Lucky's bet, but that the outcome of the bet is not, the (Lucky-bets-all-tails&Fred-flips-all-tails)-world w_{18} comes out with more causally independent imperfect match than does the (Lucky-bets-all-tails&Fred-flips-a-random-sequence)-world w_{19}. As with Hawthorne's compulsive, given that there is no distancing induced by quasi-miracles, the quasi-miracle in w_{18} wreaks no havoc. And so had Lucky bet all tails, he would have won.

Thus it seems that L plus 1c–4c solves all the problem cases here. Perhaps further problem cases remain. Perhaps there is a better way to invoke causal independence. My point is only to illustrate how causal independence proves useful.[16]

4. *Conceptual circularity*

I have recommended using causal independence in assessing standard counterfactuals. But this sort of recommendation threatens *circularity*, given that many leading approaches to causation (including Lewis 1973c, 1986c and 2000) invoke counterfactuals.

One might simply reject counterfactual accounts of causation.[17] But *need* one? Might one adopt *both* a causal independence account of coun-

[15] Or at least, to our *intuitions* the outcome of the flip is causally independent of Lucky's bet. Though in fact, there will inevitably be subtle causal influences here – perturbations in photons and sound-waves, etc., will disturb the coin's trajectory. As we awaken to the subtle causal influence here, we tend to reverse our original intuitions. Edgington notes the same point: '[T]he betting story is sensitive to whether my saying "Heads" might have influenced the manner in which the coin was tossed' (2004: 27). This is further confirmation that our intuitions are driven by causal independence intuitions.

[16] Paul Noordhof (2004: 193) has suggested invoking *probabilistic independence*. This may help with some of the cases, but not with all. In Morgenbesser's coin, for instance, the outcome of the bet remains probabilistically *in*dependent of Lucky's bet (stuck at 50–50). In general, where causal and probabilistic independence diverge, our intuitions seem to track causal independence. Perhaps there is some subtle way to 'fake' a causal independence proviso. But I see no way to do this.

[17] Thus, Edgington notes that: 'As Lewis wants to explain causal dependence and independence in terms of counterfactuals, this [need to appeal to causal independence] is a problem for him' (2004: 20). She suggests the following solution: 'give up on the idea of explaining causation in terms of counterfactuals ...' (2004: 21) And Barker writes: 'I suggest that as the CAT theory assumes that causation is conceptually prior to counterfactuals, its success in explicating PCFs, given that

terfactuals, *and* a counterfactual account of causation? Is the resulting circularity *problematic*?

Ontologically speaking, I see nothing problematic here. The truth about both counterfactuals and causality still *supervenes* on the arrangement of events. Or at least, nothing here contradicts that. The causal and counterfactual facts can still, for instance, be regarded as 'co-supervenient' upon a Humean base.

If there were a problem, it would be a *conceptual* problem. One would lose *linear definability* – no ordered chain of definitions could wind from the Humean base up through the conceptual superstructure. But perhaps linear definability was never in the offing. *Because concepts do not have definitions*. At best one can provide *rough glosses*. One can give an informative sketch of causation via counterfactual dependence, and an informative sketch of standard counterfactuals via causal independence. One cannot and need not do more.

Perhaps such circularity is only to be expected, from messy conceptions of a Humean world.[18]

University of Massachusetts-Amherst
Amherst, MA 01003, USA
schaffer@philos.umass.edu

References

Barker, S. 1999. Counterfactuals, probabilistic counterfactuals and causation. *Mind* 108: 427–69.

Bennett, J. 2001. On forward and backward counterfactual conditionals. In *Reality and Humean Supervenience: Essays on the Philosophy of David Lewis*, ed. G. Preyer and F. Siebelt, 177–202. Rowman and Littlefield: Maryland.

Edgington, D. 2004. Counterfactuals and the benefit of hindsight. In *Cause and Chance: Causation in an Indeterministic World*, ed. P. Dowe and P. Noordhof, 12–27. Routledge: London.

Elga, A. 2001. Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science* 68S: S313–24.

Fine, K. 1975. Critical notice: Counterfactuals. *Mind* 84: 451–58.

Hawthorne, J. Forthcoming. Chance and counterfactuals. *Philosophy and Phenomenological Research*.

Kvart, I. 1986. *A Theory of Counterfactuals*. Hackett Publishing: Indianapolis.

Lewis, D. 1973a. *Counterfactuals*. Basil Blackwell: Oxford.

---

other accounts fail, undercuts the project of a counterfactual analysis of causation' (1999: 428).

[18] Thanks to Troy Cross, Jamie Dreier, Adam Elga, Barry Loewer and Ryan Wasserman. This paper emerged from my comments on Wasserman's 'The future similarity objection revisited', given at the Eastern APA, Washington, 30/12/03.

Lewis, D. 1973b. Counterfactuals and comparative possibility. *Journal of Philosophical Logic* 2: 418–46.
Lewis, D. 1973c. Causation. *The Journal of Philosophy* 70: 556–67.
Lewis, D. 1979. Counterfactual dependence and time's arrow. *Nous* 13: 455–76.
Lewis, D. 1986a. *On the Plurality of Worlds*. Basil Blackwell: Oxford.
Lewis, D. 1986b. Postscripts to 'Counterfactual dependence and time's arrow'. In his *Philosophical Papers: Volume II*, 52–66. Oxford University Press: Oxford.
Lewis, D. 1986c. Postscripts to 'Causation'. In his *Philosophical Papers: Volume II*, 172–213. Oxford University Press: Oxford.
Lewis, D. 2000. Causation as influence. *The Journal of Philosophy* 97: 182–98.
Noordhof, P. 2004. Prospects for a counterfactual theory of causation. In *Cause and Chance: Causation in an Indeterministic World*, ed. P. Dowe and P. Noordhof, 188–201. Routledge: London.
Slote, M. 1978. Time in counterfactuals. *Philosophical Review* 87: 3–27.
Tooley, M. 2003. The Stalnaker-Lewis approach to counterfactuals. *The Journal of Philosophy* 100: 371–77.
Wasserman, R. Unpublished. The future similarity objection revisited.