Counterfactuals and comparative similarity

Jeremy Goodman

Draft of December 2, 2016

Abstract

An analysis of counterfactuals in terms of the comparative similarity of possible worlds is widely attributed to Lewis (1973). In this note I show that this attribution is mistaken, and argue that the view widely misattributed to Lewis is untenable.

Consider the following characterizations of the theory advanced by David Lewis in *Counterfactuals* (1973):

> The book is centred upon an analysis of counterfactuals in terms of possible worlds. The first chapter presents the analysis: the counterfactual 'if it were the case that ϕ then it would be the case that ψ' – written '$\phi \boxdot\!\!\rightarrow \psi$' – is true either vacuously or non-vacuously; vacuously if no ϕ-world is 'entertainable'; non-vacuously if, within some degree of similarity to the actual world, some possible world is a ϕ-world but none is a $\phi\&\sim\psi$-world. [...] A later formulation is in terms of the three place relation $\leq_i$ of comparative similarity among possible worlds: '$j \leq_i k$' is read 'j is as similar to i as k'. (Fine, 1975, p457)

> Lewis's theory states truth conditions for conditionals in terms of a three-place comparative similarity relation [...] $C_i(j, k)$ mean[ing] that j is more similar to i than k is to i. (Stalnaker, 1984, p133)

> [F]or Lewis, $A > C$ is true at a world w just in case some world at which $A\&C$ is true is closer or more similar to w than is any world at which $A\&\sim C$ holds. Like Stalnaker, Lewis appealed to an intuitive, pre-analytical notion of overall similarity of worlds, much like overall similarity of cities or of planets. (Lycan, 2001, p49)

> Lewis offered a nice graphic way of thinking about this. He proposed that we think of similarity between worlds as a kind of metric, with the worlds arranged in some large-dimensional space, and more similar worlds being closer to each other than more dissimilar worlds. (Weatherson, 2014)

Such quotations could be multiplied,[1] and they fit with how Lewis himself later described his view:

[1] For example, the other entry in the *Stanford Encyclopedia of Philosophy* to characterize Lewis's view says

> The central notion of a possible world semantics for counterfactuals is a relation of comparative similarity between worlds (Lewis 1973). One world is said to be closer to actuality than another if the first resembles the actual world more than the second does. (Menzies, 2014)

> A counterfactual "If it were that A, then it would be that C" is (nonvacuously) true if and only if some (accessible) world where both A and C are true is more similar to our actual world, overall, than is any world where A is true but C is false. (Lewis, 1979, p465)

Lewis goes on to liken the "overall similarity of worlds" to relations of similarity in general, and writes that the above "[a]nalysis [...] (plus some simple observations about the formal character of comparative similarity) is about all that can be said in full generality about counterfactuals. While not devoid of testable content – it settles some questions of logic...".

In fact, the analysis of counterfactuals suggested in the above quotations, together with some uncontroversial facts about the "formal character of comparative similarity", makes predictions that do not follow from Lewis's official analysis in *Counterfactuals*. Although Lewis himself discusses this point in section 2.4 of the book, this discussion seems to have gone almost completely unnoticed – indeed, as the last quotation shows, Lewis was not careful about this point himself. This slip is not a mere exegetical curiosity, since the theory attributed to Lewis in the above quotations is both widely endorsed and subject to counterexamples to which Lewis's official view is not vulnerable.

1 Two analyses distinguished

According to the view attributed to Lewis in the above quotations, for any distinct worlds v and u, the counterfactual *had either v been actualized or u been actualized, v would have been actualized* is true at a world w just in case v is more similar to w than u is to w. We can formalize this claim as follows, letting $\mathcal{A}$ abbreviate '...is actualized', $x_1y_1 \ll x_2y_2$ abbreviate 'x_1 is more similar to y_1 than x_2 is to y_2', $\boxdot\!\!\rightarrow$ abbreviate the counterfactual conditional, and 'at w' abbreviate '$\mathcal{A}w \boxdot\!\!\rightarrow \ldots$':

SIMILARITY: $v \neq u \rightarrow ((\text{at } w : (\mathcal{A}v \vee \mathcal{A}u) \boxdot\!\!\rightarrow \mathcal{A}v) \leftrightarrow vw \ll uw)$

Now, if we know anything at all about relations of comparative similarity, we know that they satisfy the following three principles:[2]

SYMMETRY: $x_1y_1 \ll x_2y_2 \rightarrow x_1y_1 \ll y_2x_2$

TRANSITIVITY: $(x_1y_1 \ll x_2y_2 \wedge x_2y_2 \ll x_3y_3) \rightarrow x_1y_1 \ll x_3y_3$

IRREFLEXIVITY: $\neg(xy \ll xy)$

Informally: The degree of similarity between two things is not relative to the order in which they are considered, the 'more similar than'-relation is transitive, and no two things are more similar to each other than they are to each other. Given these principles, SIMILARITY entails that for no three worlds x, y, and z are all three of the following the case:

---

and in his influential book on conditionals Jonathan Bennett glosses Lewis's analysis as follows

> A > C = C obtains at every member of some class W of A-worlds such that every member of W is *more like* the actual world than is any A-world that is not in W. (Bennett, 2003, p166, emphasis original)

[2]See Williamson (1988); other than Lewis (1973, section 2.4), his is the only discussion I know of that distinguishes the two views under discussion.

(i) at $x : (\mathcal{A}y \vee \mathcal{A}z) \boxdot\!\!\rightarrow \mathcal{A}y$

(ii) at $y : (\mathcal{A}z \vee \mathcal{A}x) \boxdot\!\!\rightarrow \mathcal{A}z$

(iii) at $z : (\mathcal{A}x \vee \mathcal{A}y) \boxdot\!\!\rightarrow \mathcal{A}x$

This is because there cannot be circles of comparative similarity: if y is more similar to x than z is to x, and z is more similar to y than x is to y, then x cannot be more similar to z than y is to z.[3]

But (i)-(iii) are consistent with the analysis of counterfactuals given by Lewis (1973). According to that analysis, each world w is associated with a total pre-order of worlds $\leq_w$ such that $v \leq_w w$ if and only if $v = w$. (A total pre-order of a class is a binary relation on that class that is reflexive, transitive, and total, in the sense that it holds in at least one direction between any two members of the class.) Let a φ-world be any world at which it is true that φ, and let $v <_w u$ abbreviate '$v \leq_w u$ and not $u \leq_w v$'. Lewis's analysis of counterfactuals is that $\varphi \boxdot\!\!\rightarrow \psi$ is true at w just in case either (i) there are no φ-worlds, or (ii) there is a φ-and-ψ-world v such that, for every φ-and-not-ψ world u, $v <_w u$.[4] Since it is perfectly compatible with this analysis that there be three worlds x, y, z such that $y <_x z$, $z <_y x$, and $x <_z y$, it is perfectly compatible with the analysis that there be three worlds satisfying (i)-(iii). Since such a situation is incompatible with SIMILARITY on any reasonable understanding of 'more similar than', it follows that there is no such understanding on which Lewis's analysis entails SIMILARITY.

Conversely, it is not clear whether the analysis of counterfactuals in terms of comparative similarity counts as a version of Lewis's official analysis – at least, not until more is said about the operative notion of similarity and the background theory of possible worlds. This is because it is not clear whether the relation *being no further from w than* has the formal properties of Lewis's $\leq_w$, where v is *no further from w than u* is just in case u is not more similar to w than v is to w. Although the relation is clearly reflexive and total, it is not obviously transitive: perhaps v_1 is more similar to w than v_2 is to w because v_1 and v_2 are very similar to each other and differ only in a respect in which v_1 matches w, yet both v_1 and v_2 are incomparable with u as regards

[3] *Proof*:

(1)	$x \neq y \wedge y \neq z \wedge x \neq z$	premise
(2)	$(\text{at } x : (\mathcal{A}y \vee \mathcal{A}z) \boxdot\!\!\rightarrow \mathcal{A}y) \leftrightarrow yx \ll zx$	(1), SIMILARITY
(3)	$(\text{at } y : (\mathcal{A}z \vee \mathcal{A}x) \boxdot\!\!\rightarrow \mathcal{A}z) \leftrightarrow zy \ll xy$	(1), SIMILARITY
(4)	$(\text{at } z : (\mathcal{A}x \vee \mathcal{A}y) \boxdot\!\!\rightarrow \mathcal{A}x) \leftrightarrow xz \ll yz$	(1), SIMILARITY
(5)	$yx \ll zx$	(2), (i)
(6)	$zy \ll xy$	(3), (ii)
(7)	$xz \ll yz$	(4), (iii)
(8)	$xy \ll zx$	(5), SYMMETRY
(9)	$yz \ll xy$	(6), SYMMETRY
(10)	$zx \ll yz$	(7), SYMMETRY
(11)	$xy \ll xy$	(8), (9), (10), TRANSITIVITY
(12)	$\bot$	(11), IRREFLEXIVITY

[4] For simplicity, we omit an accessibility relation.

similarity to w, since v_1 and v_2 differ radically from u, and hence neither is u more similar to w than v_2 is to w nor is v_1 more similar to w than u is to w – if so, transitivity fails. Nor does the relation obviously satisfy Lewis's 'strong centering' requirement that w be the least element under it, since it is not obvious that w must be more similar to itself than it is to any other world v – if it isn't, then the relation violates strong centering.

Interestingly, the difference between Lewis's official view and the comparative-similarity-theoretic view with which it tends to be conflated makes no difference to the propositional logic of counterfactuals. This is because any distance metric on worlds straightforwardly determines both a comparative similarity relation on worlds and a corresponding three-place relation $\leq$, and Schlechta and Makinson (1994) have shown that the propositional logic of counterfactuals determined by Lewisian models based on relations generated from distance metrics in this way is no stronger than Lewis's logic.[5] It follows that the conjunction of (i)-(iii) is satisfiable in models based on comparative similarity relations, but only on non-intended interpretations in which formulas of the form $\mathcal{A}w$ can be true at more than one index in the model.[6]

2 Against the comparative similarity analysis

I will now argue against the analysis of counterfactuals in terms of the comparative similarity of possible worlds by arguing that there are triples of worlds satisfying (i)-(iii). Consider three worlds x, y, and z and their respective laws of nature L_x, L_y, and L_z. At y there is only a 'small, localized, simple' violation of L_x, at z there is only a 'small, localized, simple' violation of L_y, and at x there is only a 'small, localized, simple' violation of L_z. By contrast, at z there are 'big, widespread, diverse' violations of L_x, at x there are 'big, widespread, diverse' violations of L_y, and at y there are 'big, widespread, diverse' violations of L_z. There is no 'spatio-temporal region [of positive volume] throughout which perfect [or even approximate] match of particular fact prevails' between any two of these worlds. Assuming there are such worlds, it is both independently plausible and predicted by Lewis (1979, p472) that (i)-(iii) are true on their most natural interpretation.

Alternatively, instead of appealing to the distinction between big and small violations of laws, we could instead appeal to the distinction between worlds at which the laws of a given world are violated and worlds at which the laws of the given world, though still true, fail to be laws. Here is a simple picture of how this might happen. Suppose there are three kinds of matter A, B and C. x contains lawfully organized A-matter, lawless disorganized B-matter, and no C-matter; y contains lawfully organized B-matter, lawless disorganized C-matter, and no A-matter; z contains lawfully organized C-matter, lawless disorganized A-matter, and no B-matter. L_x, L_y and L_z respectively concern only A-matter, B-matter, and C-matter, but (we may suppose) do not entail that there *is* any such matter, only that any such matter there may be behaves in a certain

[5]The proof of this result is non-trivial and appears not to be well known, as the same question was posed (inconclusively) by Aiello and van Benthem (2002).

[6]By letting $\mathcal{A}x$, $\mathcal{A}y$, and $\mathcal{A}z$ each be true at infinitely many indices we can validate (i)-(iii) using the following model in which $\leq$ is determined by a distance metric on indices: let $[\![\mathcal{A}x]\!] = \{1, \frac{1}{4}, \frac{1}{7}, \dots\}$, $[\![\mathcal{A}y]\!] = \{\frac{1}{2}, \frac{1}{5}, \frac{1}{8}, \dots\}$, and $[\![\mathcal{A}z]\!] = \{\frac{1}{3}, \frac{1}{6}, \frac{1}{9}, \dots\}$, and use the natural Euclidean metric on indices.

organized way. Such a case instantiates the structure described above. It trades on the idea that, although it would be hard to violate the laws, it wouldn't be as hard for violated generalizations to have been laws.

In *Counterfactuals* Lewis gives an argument in a similar spirit:

> [The assumption] that the similarity of i to j equals the similarity of j to i [...] implies a constraint on similarity orderings derived from that measure [of similarity]: if $j <_i k$ and $k <_j i$, then $j <_k i$. But that constraint would be unjustified if we suppose that the facts about a world i help to determine which respects of similarity and dissimilarity are important in comparing other worlds in respect of similarity to the world i. The colors of things are moderately important at our world, so similarities and dissimilarities in respect of color contribute with moderate weight to the similarity or dissimilarity of other worlds to ours. But there are worlds where colors are much more important than they are at ours; for instance, worlds where the colors of things figure in fundamental physical laws. There are other worlds where colors are much less important than they are at ours; for instance, worlds where the colors of things are random and constantly changing. Similarities or dissimilarities in color will contribute with more or less weight to the similarity or dissimilarity of a world to one of those worlds where color is more important or less important. Thus it can happen that j is more similar than k to i in the respects of comparison that are important at i; k is more similar than i to j in the respects of comparison that are important at j; yet i is more similar than j to k in the respects of comparison that are important at k. (Lewis, 1973, p51)

One problem with Lewis's example is that colors probably couldn't really have figured in fundamental physical laws, and it isn't immediately obvious how to repair the example. Setting aside that worry, though, here is one way of fleshing out the idea. Let i be a world in which colors figure in fundamental laws, j be the actual world, and k be a world in which there is no regularity in how things are colored. We assume that similarity in chromatic respects is all that matters for $\leq_i$, that similarity in non-chromatic respects is all that matters for $\leq_k$, and that similarity in both respects matters to some degree for $\leq_j$. If degree of regularity is anything to go by, j is more similar to i in chromatic respects than k is to i, and hence $j \leq_i k$. Assuming (unrealistically, but for the sake of argument) that colors can vary independently from non-chromatic properties, we can stipulate that i and k are alike in all non-chromatic respects, but differ from j in such respects. It follows that $i \leq_k j$. Finally, we stipulate that j is more similar to k in chromatic respects than it is to i. Since there is nothing non-chromatic to distinguish i and k from j, the chromatic difference dominates and $k \leq_j i$, completing our counterexample.

As the example demonstrates, there is nothing wrong with having a function from worlds w to comparative similarity relations $\ll_w$ such that $v \leq_w u$ just in case $vw \ll_w uw$ – Lewis (1979) in effect does exactly that. But no single comparative similarity relation can do the job of Lewis's $\leq$, since the symmetry inherent in comparative similarity relations renders them incapable of encoding all of the asymmetries there are in the counterfactual structure of modal space.

References

Marco Aiello and Johan van Benthem. A modal walk through space. *Journal of Applied Non-Classical Logics*, 12(3-4):319–363, 2002.

Jonathan Bennett. *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press, 2003.

Kit Fine. Review of David Lewis's *Counterfactuals*. *Mind*, 84:451–8, 1975.

Kit Fine. Counterfactuals without possible worlds. *The Journal of Philosophy*, 109:221–226, 2012.

David Lewis. *Counterfactuals*. Oxford: Basil Blackwell, 1973.

David Lewis. Counterfactual dependence and time's arrow. *Noûs*, 13(4):455–76, 1979.

William Lycan. *Real Conditionals*. Oxford: Oxford University Press, 2001.

Peter Menzies. Counterfactual theories of causation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, volume Spring 2014. `http://plato.stanford.edu/archives/spr2014/entries/causation-counterfactual/`, 2014.

Karl Schlechta and David Makinson. Local and global metrics for the semantics of counterfactual conditionals. *Journal of Applied Non-Classical Logics*, 4(2): 129–140, 1994.

Robert Stalnaker. *Inquiry*. Cambridge, MA: MIT Press, 1984.

Brian Weatherson. David lewis. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/archives/win2014/entries/david-lewis/`, winter 2014 edition, 2014.

Timothy Williamson. First-order logics for comparative similarity. *Notre Dame Journal of Formal Logic*, 29(4):457–81, 1988.