SIMILARITY IS A BAD GUIDE TO COUNTERFACTUAL TRUTH

ABSTRACT. The most popular theory of how to evaluate counterfactuals is to use the Lewis-Stalnaker logic together with some reasonably tractable refinement of our ordinary notions of similarity. This approach is misguided because for some ordinary counterfactuals, irrelevant possible worlds end up determining the counterfactuals' truth values. This undermines some of the support for the Lewis-Stalnaker logic, e.g. the failure of antecedent-strengthening.

The Lewis-Stalnaker logic systems [2] account for a wide range of intuitions about counterfactual inferences, even accommodating competing theories of counterfactuals like Nelson Goodman's covering law account [5]. This notable success encouraged speculation that some refinement of our ordinary intuitions about similarity informs the truth conditions of counterfactuals dealing with physical processes. The most famous is David Lewis' [3] theory of an overall similarity relation based on a priority ranking of several different respects of similarity. A number of counterexamples have been produced showing Lewis' theory produces the wrong truth values in certain cases, but the problems with theories of overall similarity are deeper than just failing to match our intuitions about particular counterfactuals. The similarity approach contains an unresolvable tension. On the one hand, the logic constrains interpretations of counterfactual sentences so that (1) the antecedent is interpreted as if it stood alone as a declarative statement, i.e. without additional contextual restrictions due to its being embedded in a counterfactual, and (2) all the context-dependence is treated as a variation of the similarity relation. This requires the similarity relation to be specially tailored to ensure for each counterfactual statement, its interpretation into a proposition sufficiently captures the meaning of the antecedent. On the other hand, to determine the truth values of particular statements, we need a relatively simple theory of overall similarity that rules out ad hoc similarity judgments.

A counterfactual is a statement about a class of possibilities partially delineated by the antecedent. ('Counterfactual' here refers to statements expressible in the form, "If *A* were

true, C would be true," without implying that the connection between A and C can be interpreted as a propositional connective.) There are two prominent competing ways to understand the relation between the antecedent and the class of possible worlds that the counterfactual is about. In covering law theories like Goodman's, the appropriate possible worlds are worlds where the antecedent is true and some (not fully specified) background conditions hold. In similarity-based approaches, the appropriate possible worlds, roughly speaking, are the A-worlds most similar to actuality, as measured by some context-dependent similarity relation.

Both approaches use these mechanisms to accommodate the fact that the meaning of a counterfactual assertion typically precludes some of the A-worlds from being relevant to its truth. Let the *relevance condition* be the claim that our theory of counterfactual truth should never let the truth value of a counterfactual depend on worlds that are definitely irrelevant given the counterfactual's meaning. Satisfying the relevance condition is more important than saving intuitions about the truth values of particular counterfactuals because our estimates about particular truths might be faulty, but any theory that violates the relevance condition implies that in some circumstances, no matter how transparent the ordinary meaning of the counterfactual is, we are mistaken about the subject of discussion.

In the similarity approach, the relevance condition is not automatically satisfied, and instead counts as a constraint on any theory of overall similarity. While we need a theory of overall similarity to guide our judgments of which counterfactuals are true; by couching overall similarity in relatively simple terms, the theory is hampered in its ability to adapt to semantic subtleties. In particular, Lewis' theory of overall similarity violates the relevance condition in several cases with the violations occurring in a way that cannot be accommodated by patching the theory because they hinge on highly contextual features of counterfactual statements that cannot be systematically codified. Hence, the prospects for a theory of overall similarity of roughly the kind Lewis' imagined is in much worse shape than current counterexamples have so far indicated.

1. Violations of the Relevance Condition

Adam Elga [1] presents an argument against Lewis' theory of overall similarity designed to show that it gets the counterfactual symmetry wrong: At 8:00, Gretta cracked open an egg onto a hot frying pan, and 5 minutes later there was cooked egg on the pan. Under some reasonable circumstances, we should argee that

1. *If Gretta hadn't cracked the egg, then at 8:05 there wouldn't have been cooked egg on the pan.*

To make it true in a deterministic context, Lewis needs to have the nearest worlds be worlds where a small miracle just before 8:00 creates a divergence such that Gretta doesn't crack the egg. Yet, (1) comes out false because there are worlds at least as similar that contain a small miracle just after 8:05, perfect match of the actual world after this miracle, and ordinary lawful physics before the miracle. These are worlds where there is cooked egg on the pan at 8:05 but where the dynamical retrodiction of the physical state going back to 8:00 is so extensively altered by the dynamical consequences of the miracle that Greta doesn't even exist. Hence, it is false that she cracks the egg. Statistical mechanics bolsters the argument by making plausible that virtually any significant molestation of a few atoms at 8:05 will imply a higher entropy state at 8:00, and such a state is overwhelmingly likely to exclude Gretta's existence.

While Elga's argument is successful, the problem is not that Lewis' theory fails to be physically sophisticated enough, but with the fact that the antecedent is construed merely as the negation of an ordinary positive claim, "Gretta cracked the egg." To see this, first note how the counterexample doesn't work if the antecedent is expressed as a positive failure:

2. *Had Gretta failed to crack the egg, there wouldn't have been cooked egg on the pan.*

Even though this is a reasonable gloss on what we ordinarily mean when we assert (1), the counterexample world Elga gives us does not have Gretta failing to crack the egg because she doesn't exist at 8:00 or before.

Second, we can construct other examples having nothing to do with statistical mechanics that violate the relevance condition by exploiting negation in the same way. Imagine a button that reliably launches the nuclear arsenal and Nixon standing nearby. Fortunately, Nixon avoids pressing the button at time t by deciding better of it. Lewis addresses the counterfactual,

3. *If Nixon had pressed the button (at time t), there would have been an apocalypse,*

by arguing that the closest A-worlds are like w_1, where a small miracle alters Nixon's brain just before t, making him decide to press the button and cause the holocaust. We can reword the same counterfactual postulation with a slightly different sentence,

4. *If Nixon hadn't avoided pressing the button (at time t), there would have been an apocalypse.*

A reasonable interpretation of the meaning of (4) has the relevant A-worlds being worlds where Nixon presses the button. However, the similarity approach forbids us from construing the sentence in this natural way. We must interpret the antecedent as the negation of "Nixon avoided pressing the button," and then hope that the overall similarity relation narrows the class of A-worlds to those where Nixon presses the button. Yet, under Lewis' theory of overall similarity there is a problematic world w_8 that is closer to actuality than w_1. By stipulation, w_8 possesses a longer stretch of perfect match up until just before t, and then has a small blast of energy inside Nixon's head very quickly frying his brain to a crisp. It preserves a bit more perfect match and involves only an arbitrarily small region for the miraculous infusion of energy. The world w_8 falsifies (4) because Nixon is not alive at t, and therefore it is not the case that he avoids pressing the button.

Lewis' theory of overall similarity is thus forced to draw fine distinctions between 'pressing the button' and 'not avoiding pressing the button' in contexts where these antecedents intuitively pick out the same class of A-worlds. It must also distinguish between 'not cracking the egg' and 'failing to crack the egg', and so on for a wide range of idioms that cannot be easily systemitized. Because the theory of overall similarity is supposed to codify in a simple way what we intuitively have in mind when we evaluate counterfactuals,

the context sensitivity one needs to successfully address the wide range of counterexamples from idiomatic subtleties is unachievable.

The upshot of these examples is that when evaluating counterfactuals, we shouldn't model antecedents as bare propositions, i.e. all the logically possible worlds where the antecedent is true. We should incorporate contextual factors directly into a restriction on the antecedents, focusing on a class of relevant antecedent-worlds. For "If Gretta hadn't cracked the egg," the relevant worlds include Gretta being alive for a least a while and then somehow not cracking the egg. At best, fans of the Stalnaker-Lewis logic can use similarity as *part* of a counterfactual's truth conditions, but must restrict the antecedent to relevant *A*-worlds. The difference is that the background conditions spelling out which *A*-worlds are relevant cannot always be made precise solely by appeal to a theoretically-based notion of similarity.

2. Strengthening the Antecedent

Once we have accepted the need to restrict the *A*-worlds beyond what can be accounted for by similarity, then some of the motivation for the Lewis-Stalnaker logic can come under attack as well. Michael McDermott [6] has questioned a key argument for the counterfactual logic, the invalidity of Antecedent Strengthening. McDermott contrasts Lewis' analysis with his preferred alternative, a 'forking' account where the relevant *A*-worlds are worlds identical to actuality before some 'forking' time and afterwards evolve indeterministically in a way that makes *A* obtain:

> Lewis based his closeness analysis on the claim that Antecedent Strengthening fails for subjunctives: for example we can accept (5) and reject (6).
>
> **5.** *If I walked on the grass, the lawn would not be ruined.*
>
> **6.** *If we all walked on the grass, the lawn would not be ruined.*
>
> I think this argument is wrong: the natural interpretations of (5) and (6) use different resolutions of vagueness. The salient fork for (5) is just my decision whether or not to walk on the grass, but for (6) the salient fork is

> a combination of several decisions. We can apply to (5) the interpretation that seems natural for (6), but on that interpretation we reject (5): 'No, not necessarily. It depends on what the others do'.

McDermott's appeal here to a "natural" interpretation of each counterfactual is comparable to my proposal that the set of relevant *A*-worlds is sometimes better given by a direct examination of the meaning of the counterfactual and not indirectly through a theory of similarity. In response to McDermott's argument, a defender of the similarity approach can object that intuitions about naturalness of interpretation are rather weak evidence, and when weighed against the range of putative benefits the similarity approach provides, McDermott's intuition can be set aside. However, my argument above shows that the similarity theorist cannot uniformly reject appeals to natural interpretations of what the relevant *A*-worlds are. Similarity alone doesn't always pick out the right *A*-worlds, assuming that similarity is not so theoretically free that it can be tailored on an ad hoc basis. Thus, McDermott's appeal to different contexts cannot be dismissed on principle, and his argument undermines a signature justification for the Lewis-Stalnaker logic.

References

[1] Adam Elga, "Statistical Mechanics and the Asymmetry of Counterfactual Dependence," *Philosophy of Science* Suppl. Vol., 2000. 3

[2] David Lewis. *Counterfactuals*. Cambridge: Harvard University Press, 1973. 1

[3] David Lewis, "Counterfactual Dependence and Time's Arrow," *Noûs* **13** (1979), 455–76, reprinted in *Philosophical Papers, Volume 2*, Oxford: Oxford University Press, 1986. 1

[4] David Lewis, "Causal Decision Theory," *Australasian Journal of Philosophy* **59** (1981), 5–30, reprinted in *Philosophical Papers, Volume 2*, Oxford: Oxford University Press, 1986.

[5] Barry Loewer, "Cotenability and Counterfactual Logics," *," The Journal of Philosophical Logic* **8** (1979), 99–115. 1

[6] Michael McDermott, ""Critical Notice of Jonathan Bennett's *A Philosophical Guide to Conditionals*," *," Australasian Journal of Philosophy* **82** 2 (2004), 341-350. 5