Check for updates

DOI: 10.1111/phpr.13138

Philosophy and Phenomenological Research

ORIGINAL ARTICLE

# Similarity accounts of counterfactuals: A reality check[1]

**Alan Hájek**

Philosophy, RSSS, Australian National University, Canberra, Australian Capital Territory, Australia

**Correspondence**: Alan Hájek
Email: alan.hajek@anu.edu.au

**Abstract**

To an unusual extent, philosophers agree that counterfactuals have truth conditions involving the most similar possible worlds where their antecedents are true, in the style of the celebrated and path-breaking Stalnaker/Lewis accounts. Roughly, these accounts say that the counterfactual *if A were the case, C would be the case* is true if and only if *at the most similar A-worlds, C is true*. I will argue that there are general structural problems with the appeals to both "the most" and "similar". I will challenge any fixation on 'the most __ worlds', however we fill in the blank with a non-trivial ordering of worlds: in ignoring worlds that are later in

[1]My foremost gratitude goes to Daniel Berntson, Brian Hedden, Alexander Kocurek, Wolfgang Schwarz, and Daniel Stoljar for ongoing discussion over years and for their trenchant and detailed comments on earlier drafts that led to multiple improvements. For very helpful comments on drafts and discussion I also thank especially Kyle Blumberg, Szymon Bogacz, Chris Bottomley, Ray Briggs, Justin D'Ambrosio, Nicholas DiBella, Stephen Finlay, Melissa Fusco, Mario Günther, Leon Leontyev, Karen Lewis, Yael Loewenstein, Barry Loewer, Sebastian Liu, Matthew Mandelkern, Cory Nichols, Daniel Nolan, Philip Pettit, Una Stojnić, Hezki Symonds, Michael Titelbaum, Barbara Vetter, Timothy Luke Williamson, James Willoughby, and an anonymous reviewer for *Philosophy and Phenomenological Research*. I am also grateful to many others, including Sharon Berry, Dorothy Edgington, Brian Garrett, Dmitri Gallow, Renée Hájek, Lloyd Humberstone, Yoaav Isaacs, Justin Khoo, Boris Kment, Hanti Lin, Neil McDonnell, Sarah Moss, Dan Munoz, Graham Oddie, Pamela Robinson, Alex Sandgren, Paolo Santorio, Robert Stalnaker, Katie Steele, and audiences at Academia Sinica, University of Arizona, Baylor University, Belgrade University, Bilkent University, Bristol University, University of Calgary, Cambridge University, Carleton University, Complutense University, University of Georgia, Konstanz University, Monash University, Rutgers University, University of Sydney, Trnava University, UCLA, University of Virginia, the Australasian Association of Philosophy Conference (ANU), the LOGOS Workshop on Conditionals at Universitat de Barcelona, and the Munich-Venice Probability Conference.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *Philosophy and Phenomenological Research* published by Wiley Periodicals LLC on behalf of Philosophy and Phenomenological Research LLC.

the ordering, it adjudicates various implausibly specific counterfactuals to be true. I will then raise foundational problems for appealing to 'similarity'—from consequents that are chancy, disjunctive antecedents, and unspecific antecedents more generally. I will also raise further problems for a number of specific proposals for understanding 'similarity'. A recurring theme will be the tension that may arise between probability and similarity considerations. I will end by arguing for a paradigm shift, replacing 'the most similar worlds' approach with one based on conditional chances.

**KEYWORDS**
Counterfactuals, similarity, Stalnaker, Lewis, chances

## 1 | SIMILARITY AS ORTHODOXY

They say that philosophers never agree—except, perhaps, on this very claim. But to an unusual extent, philosophers agree that counterfactuals have truth conditions involving the most similar possible worlds where their antecedents are true, in the style of the celebrated and path-breaking Stalnaker/Lewis accounts. Roughly, these accounts say that the counterfactual

> *if A were the case, C would be the case*

is true if and only if

> *at the most similar A-worlds, C is true.*

Most philosophers agree with this, as I used to. But now I beg to differ—I will argue against this entire approach.

Nichols (2021, 10535) observes that this approach

> has been, by far, the predominant paradigm within the counterfactuals literature, in both philosophy and linguistics, for around 50 years. Its status as orthodoxy is so firmly established that it is often appealed to without argument by authors working on related subjects …

He cites works by several authors in various different areas. We could continue almost indefinitely the list of appeals to this approach across philosophy. To be sure, some alternative approaches to counterfactuals have also been developed.[2] There have also been a number of critics of the

[2] They include structural equations approaches (e.g. Briggs, 2012), truthmaker approaches (e.g. Fine, 2012), and chance approaches that I will discuss later.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

most-similar-worlds paradigm.[3] They target the appeal to 'similarity', and mainly the particular details of its implementation. This is all to the good, and I will also target these things; but I will go further, pointing out general structural problems with any such appeal that run deep, and also targeting the fixation on the *most* similar worlds. I will argue for an entire paradigm shift.

In §2 I will start with Robert Stalnaker's original version of such an account, quickly rehearse some old objections to it, and suggest some new ones. This will prompt David Lewis's alternative account in §3. Still, the spirit is very much the same. In §4 I will challenge fixating on 'the most __ worlds', however we fill in the blank with an ordering of worlds: in ignoring worlds that are later in the ordering, it adjudicates various implausibly specific counterfactuals to be true. In particular, in §5 I will challenge filling in the blank with a *similarity* ordering, in which worlds are ordered by their *resemblance* (in some suitable sense) to the actual world, or more broadly to a given base world. At a first stab, in §5.1 we will understand similarity to be an intuitive resemblance relation that the folk would readily understand and recognize. This will quickly lead to some well-known counterexamples, and I will offer more. §5.2 will discuss 'similarity' regarded as a 'black box', an unspecified ordering. We will look in §5.3 at Lewis's more technical specification of the respects of resemblance that matter to his account, and further refinements by him and by J. Robert G. Williams to accommodate indeterministic worlds in §5.4 and §5.5. Doing so will display a heuristic for generating particularly problematic counterexamples, in which the most __ worlds are the most *improbable*, given the counterfactuals' antecedents. More generally, a recurring theme will be the tension that may arise between probability and similarity considerations.

I intend all this to dispel any thought that we are just quibbling about the details—that 'the most similar worlds' accounts are roughly on the right track, and we just need to spell out better what matters to similarity. In raising foundational problems that strike these accounts more deeply, I aim to convince you that they will not be solved by a clever tweak here or a tinker there. I will end by arguing for a replacement of 'the most similar worlds' approach—the paradigm shift that I advocate—based on conditional chances.

## 2 | STALNAKER'S ACCOUNT

Stalnaker and Lewis motivate their accounts, present them, and then showcase their fruits; these fruits may be regarded then as indirectly supporting the accounts. And there is no doubt that they have been fruitful—though I will later argue their fruits can be gained in a better way.

Stalnaker (1968) firstly motivates his account by Ramsey's famous prescription for evaluating a conditional. Ramsey (1928/1990) wrote:

> If two people are arguing 'if $p$ will $q$?' and both are in doubt as to $p$, they are adding $p$ hypothetically to their stock of knowledge and arguing on that basis about $q$. We can say that they are fixing their degrees of belief in $q$ given $p$.

Stalnaker offers this understanding of the prescription:

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether the consequent is then true. (44)

[3] They include Nichols, Fine (1975, 2012), Tichý (1976), Slote (1978), Morreau (2010), Kroedel and Huber (2013), and Santorio (2019).

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Writing during a golden age of possible worlds semantics for modal logic, Stalnaker goes on to say that

> the problem is to make the transition from belief conditions to truth conditions … The concept of a *possible world* is just what we need to make this transition, since a possible world is the ontological analogue of a stock of hypothetical beliefs. (44–45)

Let me also emphasize how the Ramsey quote ends with a claim about *conditional probabilities*. The ontological analogue of a conditional credence is a *conditional chance*. So I take Ramsey also to motivate an account of counterfactuals in terms of conditional chances. I will return to this point at the end when I offer such an account.

Stalnaker goes on immediately to give "a first approximation" of his truth conditions for "If *A*, then *B*". His truth conditions are the same for counterfactuals and indicative conditionals; they differ only in the "selection function" that chooses the minimally-different antecedent-world, and this is solely a matter of pragmatics. It is striking that most-similar-world accounts are not as popular for indicative conditionals as for counterfactuals. (For example, Lewis, 1986b; Jackson, 1987; and Williamson, 2020 among others advocate the material conditional analysis of indicatives.) Yet there should be a strong presumption in favor of giving as unified an account of them as possible.[4] After all, 'if' seems to mean much the same thing in both cases. Offhand, most-similar-worlds accounts should be equally popular for indicatives and for counterfactuals: either more popular for the former, or less popular for the latter.

Stalnaker then states his truth conditions more formally in terms of a selection function that takes a world *w* and a proposition *A* as inputs, and outputs another world—the 'closest' world to *w* at which *A* is true. This provides an abstract semantics, in which the selection function could be interpreted in various ways, or not at all. But he explicitly gives the following gloss: "the selection is based on an ordering of possible worlds with respect to their **resemblance** to the base world" (46). (My bolding here and in other quotes.) And in his (1984, 141) he writes:

> Even if we take the selection function as the basic primitive semantic determinant in the analysis of conditionals, we still must rely on some more or less independently understood notion of **similarity** or closeness of worlds to describe the intuitive basis on which the selection is made. The intuitive idea is something like this: the function selects a possible world in which the antecedent is true but which otherwise is as much **like** the actual world, in relevant respects, as possible.

We can summarize Stalnaker's truth conditions as follows:

> 'if it were the case that *A*, it would be the case that *C*' is (non-vacuously) true at *w*
>
> iff
>
> *C* is true at *the A-world most similar* to *w*. (If *A* is impossible, then the counterfactual is vacuously true.)

And so the *most-similar-worlds* approach to counterfactuals was born.[5]

[4] Unified accounts include those of Weatherson (2001), Nolan (2003), Günther and Sisti (2022), and Khoo (2022).

[5] Or reborn. Todd's (1964) similarity account predates those of Stalnaker and Lewis, yet it is not nearly as well known.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

## 2.1 | 'The'

The "the" in Stalnaker's account carries his assumption that for any *A*, there is exactly *one* most similar ("closest") *A*-world. Lewis (1973c) challenges this assumption on both sides:

(i) He argues that there may be *no* closest *A*-worlds. 'If I were over 7 feet tall, then …' How tall would I be? 7 feet 1 inch? 7 feet ½ inch would be nearer to my actual height. 7 feet 1/3 inch would be nearer still … Suppose that the similarity relation favors closeness to my actual height (other things being equal). We have an infinite sequence of ever-closer 'over 7 feet tall' worlds, with none closest. This challenges the 'limit assumption' that for any *A*, there are closest-*A* worlds. Remember this example: it will come back to haunt Lewis later.

(ii) There may be *more than one* closest *A*-world. 'If Bizet and Verdi were compatriots, then they would both be …' French? Italian? It may be that nothing favors one nationality over the other. So there are at least *two* equally good candidates for 'the' closest world in which they are compatriots. This is a problem of *unspecificity*: there is no specific nationality they would both be. In a slogan: unspecificity undermines 'would'! Said more carefully: an antecedent's unspecificity undermines a 'would' with an excessively specific consequent. Remember this example also: we will later see that Lewis's own account judges various implausibly specific counterfactuals to be true.

We may strengthen this concern. Consider the greatest possible mismatch between the specificity of the antecedent and that of the consequent: a maximally unspecific (false) antecedent, and a maximally specific consequent. "If things had been different from how they actually are, then world *w** would be the case". There is a specific *w** that renders this true according to Stalnaker. But given how open the antecedent leaves things, it seems that the consequent closes things off too decisively. In virtue of what? Of course, if things had been different from how they actually are, they would be some particular way or other. But it is not clear that there is a particular way such that things would be *that* way (and no other).

Let me add that there are also problems of *indeterminism*. Suppose for now that coin tossing is indeterministic, and for that reason the outcome of a coin toss is chancy. I did not toss the fair coin in my pocket. 'If I had tossed the coin, then it would have landed …' Heads? Tails? By Stalnaker's (1968) lights, one of the answers must be right, and the other wrong. But there is reason to balk at each answer. Chance undermines 'would'!

Again, we may strengthen this problem. Consider an indeterministic process with an uncountably infinite state space that is never realised. Suppose that radioactive decay is indeterministic. Here is a tritium atom that never decayed. 'If the atom had decayed, it would have done so exactly at *t*'. According to Stalnaker, this is true for the decay time *t* in the closest decay-world. But all the more there is reason to balk at such a precise resolution of this process.

Some philosophers insist that (non-trivial) chances are incompatible with determinism—e.g. Popper (1959), Lewis (1980), and Schaffer (2007). However, there are a number of *compatibilists* about chance who believe there can be chances in a deterministic world—e.g., Levi (1990), Albert (2000), Loewer (2001, 2020), Eagle (2011), Ismael (2011), Strevens (2003, 2013), and Hoefer (2019), and I want to remain open to such compatibilism. Statistical mechanics, a probabilistic theory

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

with deterministic dynamics, is its poster child. Chanciness may be regarded as generated by unspecificity regarding initial conditions. Edgington (2004, p.14) writes:

> Even if we do live in a deterministic world, we do not live in a crudely deterministic world. Our ordinary run-of-the-mill antecedents are not normally specific enough to be fed into deterministic laws. Even if coin-tossing is a deterministic process, no deterministic conclusion comes from the counterfactual supposition that you had tossed the coin, but only from a supposition of how *exactly down to the minutest detail* you tossed it.

So we should still balk at each putative answer to how the coin would have landed if I had tossed it, even under determinism—this time because of the unspecificity of its antecedent making the outcome chancy in a compatibilist sense.[6] Stalnaker himself (1981) offers a semantics in which the selection of the closest $A$-world is made arbitrarily; it then supervaluates over all such arbitrary selections, so that counterfactuals such as these will come out indeterminate. This moves his account closer to Lewis's insofar as they can agree that the relevant counterfactuals are *not true*, but 'indeterminate' is still not the same verdict as Lewis's 'false' in such cases.

This brings us to Lewis's account. I will defer my concerns about "most" and "similar" in Stalnaker's account, since they will apply to Lewis's also. He apparently solves the "the" problems for Stalnaker's account that he raised.

## 3 | LEWIS'S ACCOUNT

The similarities between Lewis's account and Stalnaker's are more striking than their differences. Here is Lewis's (1973a, 1973b, 1973c):

> 'if it were the case that $A$, it would be the case that $C$' is (non-vacuously) true at $w$
>
> iff
>
> some $A$ & $C$ world is more similar to $w$ than any $A$ & $\neg C$ world. (If $A$ is impossible, then the counterfactual is vacuously true.)

Lewis appeals to *comparative* similarity of worlds: "… *more* similar …". We saw that Stalnaker appeals to a *superlative*: "… *most* similar …". But we may also regard Lewis's account that way. Symbolize the counterfactual with antecedent $A$ and consequent $C$ as '$A \mathbin{\square\!\!\rightarrow} C$'. We may follow his reformulation of his account (1973b, p.49, with minor notational changes):

> $A \mathbin{\square\!\!\rightarrow} C$ is (non-vacuously) true at a world $w$
>
> if and only if
>
> there is an $A$-world $k$ such that, for any world $j$, if $j$ is at least as similar to $w$ as $k$ is, then $A \supset C$ holds at $j$.

[6] To be sure, there have been a number of defenses of Stalnaker's assumption that there is always a closest $A$-world. More recent ones include Swanson (2012), Santorio (2017), and Mandelkern (2019). I do not have the space here to engage with these; I don't need to because my main concerns lay elsewhere.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

More simply:

> $A \mathbin{\square\!\!\rightarrow} C$ is (non-vacuously) true at $w$
>
> if and only if
>
> there is an $A$-world $k$ such that $C$ holds at every $A$-world at least as similar to $w$ as $k$.

Having identified a candidate for $k$ over which the reformulation existentially quantifies, we may regard all $A$-worlds at least *that* similar to $w$ as "the most similar $A$-worlds". Even though they are not all equally similar, they all count as sufficiently similar. English permits this flexible interpretation of "most". The internet is full of "the most __" lists, *strictly ordered*. What matters is that anything on such a list is more __ than anything outside the list. And so it is with Lewis's 'more similar' account.

In any case, as we will see next, Lewis should accept the limit assumption, in which case there is always a set of maximally close $A$-worlds to a given world $w$.

## 3.1 | The limit assumption

We saw that Lewis officially rejects the assumption that there is always a most similar $A$-world. But this creates even worse problems for him.

Recall the problem that he raises: make the antecedent 'I am over 7 feet tall', and suppose that the similarity relation favors closeness to my actual height: the closer, the more similar (other things equal). Notice that he does not complete the example: he does not state an entire counterfactual that is supposed to be a counterexample to the limit assumption. In any case, the example leads to disaster for Lewis. Pollock (1976), Fine (2012) and others have made versions of the point. I offer the following as my sharpest way of putting the point, and this way will be useful later.

Here's an intuitively valid argument form—and both Stalnaker's and Lewis's accounts judge it to be valid:

*Agglomeration*
$A \mathbin{\square\!\!\rightarrow} C_1$
$\underline{A \mathbin{\square\!\!\rightarrow} C_2}$
$\therefore A \mathbin{\square\!\!\rightarrow} (C_1 \,\&\, C_2)$

It is a rule of conjunction introduction for the consequents of counterfactuals with a fixed antecedent. I submit that the argument form remains valid in the countably infinite case—*countable agglomeration*: continue the sequence of premises indefinitely and put their infinite conjunction in the conclusion's consequent. But then we have by Lewis's lights:

> if I were over 7 feet tall, I would be less than 7 feet 1 inch;
> if I were over 7 feet tall, I would be less than 7 feet 1/2 inch;
> if I were over 7 feet tall, I would be less than 7 feet 1/3 inch;
> …

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

By countable agglomeration (and equivalence in the consequent):

> if I were over 7 feet tall, I would be less than 7 feet 1/n inch, for all n. (1)

Assume that heights can be represented by real numbers. Then we should all think that

> if I were over 7 feet tall, I would be 7 feet $\varepsilon$ inch, for some real $\varepsilon > 0$. (2)

By (finite) agglomeration on (1) and (2):

> if I were over 7 feet tall, I would be less than 7 feet 1/n inch, for all n, and
>
> I would be 7 feet $\varepsilon$ inch, for some real $\varepsilon > 0$.

That is,

> if I were over 7 feet tall, CONTRADICTION.

The only way that this can be true (vacuously) is for it to be impossible for me to be over 7 feet tall. This defeats the point of the example, it is implausible in any case, and it renders trivially true all counterfactuals with the same antecedent. If I were over 7 feet tall, I would bump my head on the sky.[7]

We don't need the example to be so far-fetched. Surely I could have been different from my actual height. But if I were taller than my actual height, we reach a contradiction in a similar way. And if I were shorter than my actual height, I would have been taller than that height minus $1/n$ of an inch, for all $n$—reaching a contradiction again.

Even without countable agglomeration, Lewis's putative counterexample to the limit assumption backfires. Let's give it the tiniest tweak: make the antecedent "I am *at least* 7 feet tall". And let me complete the example: "If I were *at least* 7 feet tall, I would be *exactly* 7 feet tall". This is implausibly specific: 7.0000 … feet tall to infinitely many decimal places?! Yet it comes out true by the very same ordering that Lewis assumed in his rejection of the limit assumption.[8] Recall how Lewis objected to Stalnaker's account favoring one specific nationality over another in 'If Bizet and Verdi were compatriots …' when there seem to be two candidates: French and Italian.

[7] Cf. Lewis (1973c, p.424).

[8] What if heights can be hyperreal-valued? Then we immediately have that one might have been within an *infinitesimal* of one's actual height—but any greater discrepancy than that would have been impossible! This is still absurd. If we assume the axiom of choice, we get a contradiction again. (Thanks here to Nicholas DiBella for this point and for what follows.) Let $L$ be the set of all lengths greater than 7 feet, including lengths with non-standard parts. (Assume L is a set.) If the Axiom of Choice is true, then $L$ can be well-ordered as $L_1, L_2, \ldots$ Maximally general (set-theoretic) Infinitary Agglomeration then yields:

> If I were greater than 7 ft tall, then I would be under $L_1$ tall.
> If I were greater than 7 ft tall, then I would be under $L_2$ tall.
> …

---

> ∴ If I were greater than 7 ft tall, then I would be under $L_1$ tall *and* under $L_2$ tall *and*…
> However, the conclusion's consequent contradicts its antecedent!

If the collection of all lengths greater than 7 feet is a proper class, then appealing to the Axiom of Global Choice yields the same problem.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Lewis is committed with this ordering to *extreme* favoritism in singling out 7.0000 … feet tall for my counterfactual height when there seem to be uncountably many candidates.[9]

So let's reject his rejection of the limit assumption—let's reinstate it. Then Lewis's account of counterfactuals looks even more like Stalnaker's:

> 'if it were the case that $A$, it would be the case that $C$' is (non-vacuously) true at $w$
>
> iff
>
> $C$ is true at *all the A-worlds most similar* to $w$.

(See Lewis's own presentation of this in 1973a, 561.) The only difference from Stalnaker's account is that Lewis's allows there to be multiple most similar $A$-worlds: Bizet and Verdi being both French/Italian, the untossed coin landing heads/tails, and so on. Both accounts now explicitly appeal to "most".[10]

Earlier I observed that Lewis's original "… more similar … " account may be regarded as a "… most similar …" account—while it does not appeal to what is true at *maximally* similar antecedent-worlds, it does appeal to what is true at *sufficiently* similar worlds. But even setting that point aside, the arguments that I will turn to next target any account that ignores certain possibilities that have positive chance. This is most easily demonstrated for "… most similar …" accounts, but my arguments may be rewritten to target Lewis's original "… more similar …" account too. The problem arises from an ordering of worlds and a selection function that takes only certain worlds to be relevant to the truth of the counterfactual, given the antecedent, when other worlds need to be considered too.[11] To streamline my presentation, I will mostly concentrate on "… most similar …" accounts, noting that the arguments generalize, and in a couple of instances I will show how.

## 4 | "MOST" (AND "MORE")

Lewis argues against a strict conditional account, mainly because of what he takes to be its unhappy verdicts on the validity and invalidity of various argument forms. Consider:

*Antecedent strengthening*

$$\frac{A \mathrel{\Box\!\!\rightarrow} C}{\therefore (A \,\&\, B) \mathrel{\Box\!\!\rightarrow} C}$$

*Transitivity*

$$\frac{A \mathrel{\Box\!\!\rightarrow} B}{}$$

[9] This is so even if we rule out heights that exceed 7 feet by too much—e.g. over 8 feet.

[10] Berntson (MS) also raises problems for restricting our attention to the *most* similar $A$-worlds, and he has been a source of inspiration to me.

[11] I am grateful to an anonymous referee for *Philosophy and Phenomenological Research* for this point.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$\underline{B \mathbin{\Box\!\!\rightarrow} C}$

$\therefore A \mathbin{\Box\!\!\rightarrow} C$

*Contraposition*

$\underline{A \mathbin{\Box\!\!\rightarrow} C}$

$\therefore \neg C \rightarrow \neg A$

These forms are valid for the strict conditional. Yet various instances of them for the counterfactual strike us as invalid—see Stalnaker (1968) and Lewis (1973b)—and they are happily judged as invalid by their accounts. We have also seen that *agglomeration* is an intuitively valid argument form, and Stalnaker and Lewis adjudicate it as such. These are thought to be strong reasons in favor of their most-similar-worlds semantics.

But 'similarity' does none of the work here. All of it is done by 'most' or 'more': fixating on either the maximal members of an ordering (Stalnaker, Lewis with the limit assumption), or all worlds up to a certain point in the ordering (original Lewis). The ordering could instead be by size, or goodness, or badness, or beauty, or ugliness, or smelliness, or __ (whatever). Indeed, it could be by *dissimilarity*.[12] For example, suppose that the most dissimilar $A$-worlds are $C_1$-worlds, and that the most dissimilar $A$-worlds are $C_2$-worlds; it follows that the most dissimilar $A$-worlds are $(C_1 \,\&\, C_2)$-worlds—agglomeration is validated. It is easy to check that the most-dissimilar-worlds semantics invalidates antecedent strengthening, transitivity, and contraposition. Of course, nobody would take seriously this semantics for other reasons, but it highlights how much work 'most' does on its own in providing the logic. The same goes for 'more'.

One might think that at least delivering these validity verdicts is good news for 'most __ worlds' accounts. But there is less good news, concerning counterfactuals with consequents that are chancy, those with disjunctive antecedents, and more generally those with unspecific antecedents.

## 4.1 | Consequents that are chancy

Any 'most __ worlds' account of a counterfactual plays favorites among its antecedent-worlds, ignoring those that are further out in its ordering. However, when a counterfactual involves a chancy process, none of its positive-chance outcomes should be ignored—or at least, none of its outcomes with *sufficient* chance should be ignored. (We might argue about what counts as *sufficient* chance, but we can surely agree that outcomes with chance ½ should not be ignored, for starters.) Problems arise for such an account, then, when the two considerations come apart. For example, consider a fair coin that was not tossed. 'If it had been tossed, it would have landed *heads*' should not come out true. This refutes any 'most __ worlds' account that favors heads-worlds over tails-worlds in its ordering, for whatever reason—size, goodness, …, or similarity.

A recipe, then, for counterexamples is to fashion chancy cases in which some outcomes are less __ than others. Especially simple cases will be ones in which a coin toss, say, discriminates

[12] If there is an infinite sequence of ever more __ $A$-worlds, we may follow Lewis more closely in the analysans: there exists an $(A \,\&\, C)$-world more __ than any $(A \,\&\, \neg C)$-world. Thanks to Wolfgang Schwarz for the 'dissimilarity' example.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

between more and less __ outcomes. We could apply this recipe to various 'most similar worlds' accounts, but here I want to emphasize its generality, and to isolate the culprit: 'most'.

Suppose that $w$ is an $A$-world such that:

(i) $w$ is not among the most __ $A$-worlds to the actual world;
(ii) if $A$ were the case, there would be some chance of $w$ obtaining.

Then the 'most __ worlds' analysis predicts via (i) that if $A$ were the case, $w$ would not obtain. But this conflicts with (ii), I submit. The chance facts leave the door open to $w$ obtaining, but the door is closed via (i). In virtue of what? Much the same point may be made for 'more __ $A$-worlds' accounts, which ignore all worlds beyond a certain point in the __-ordering, while the chance facts entail that they should not be ignored.

There is a striking disanalogy between chancy processes that actually play out and merely counterfactual chancy processes. Consider an actual coin toss. There is no mystery about its result: we can film it, multiple people can witness it, and so on. Not so a merely hypothetical coin toss. The coin in my pocket was never tossed; it is much more mysterious how there could be a fact of the matter of how it *would* have landed *if* it had been tossed. To be sure, some philosophers posit *counterfacts* that serve as truthmakers for such counterfactuals. They are primitive modal facts that do not supervene on non-modal facts—see Hawthorne (2005), Bradley (2012), Schulz (2017), and Stefánsson (2018). But this only highlights the mystery: these *primitive* modal facts are ungrounded, and nothing more is said about them. And the resulting ontology is profligate. It is qualitatively unparsimonious: it introduces a new kind of entity, primitive counterfacts. And it is spectacularly quantitatively unparsimonious: it posits *many* of them. For every possible counterfactual antecedent, there is a counterfact of how it *would* be realized. And presumably for every counterfact, which could have been otherwise, there is a higher-order counterfact about what the alternative counterfact would have been—an infinite regress. (See Hájek, 2021 for further discussion.)

To avoid these concerns, ___ would have to be such that every world that would have a positive chance of obtaining were $A$ the case is also among the most ___ $A$-worlds. But then we are well on the way to analyzing counterfactuals in terms of chances—as I believe we should (see §6). (Arguably, the most ___ $A$-worlds do not include worlds that would have *no* chance of obtaining were $A$ the case.)[13]

## 4.2 | Disjunctive antecedents

What do you think of this sentence: 'If we were in Rome or Baghdad, we would be in Italy'? I say that's false, and so does nearly everyone I have asked. The reasoning is simple: *Baghdad is not in Italy*. Why does that matter? We mentally test the simpler counterfactuals:

> 'If we were in Rome, we would be in Italy'. (Tick!)
>
> 'If we were in Baghdad, we would be in Italy'. (Cross!)

[13] I am grateful to Alexander Kocurek for formulating the argument this way.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Since the latter counterfactual is false, so must be the original disjunctive-antecedent counterfactual, which apparently implies it.

This is an instance of the well-known inference rule:

*Simplification of Disjunctive Antecedents* (SDA):

$$\frac{(A \vee B) \mathbin{\Box\!\!\rightarrow} C}{\therefore (A \mathbin{\Box\!\!\rightarrow} C) \,\&\, (B \mathbin{\Box\!\!\rightarrow} C)}$$

SDA was proposed by Nute (1975), and it is upheld by a number of authors, including Fine (1975), Ellis et al. (1977), Briggs (2012), and Santorio (2018). Others have noticed the disjunctive-antecedents problem for the Stalnaker/Lewis-style most-similar-worlds accounts—e.g. Fine (1975) and Ellis et al. (1977); see Nute and Cross (2001) for an overview. In our example, we may easily imagine the most similar world in which we are in Rome or Baghdad is one in which we are in Rome: we are more disposed to go to Rome, it is easier to travel to, etc.; and Rome *is* in Italy. I want to stress that disjunctive antecedents (at least where the disjunction is understood classically) threaten to be a problem for non-trivial '*most*-__-worlds' accounts, however we fill in the blank—similar, beautiful, ugly, smelly, etc… Such an account only looks at the 'front row' of $(A \vee B)$-worlds, as ordered by whatever. Disjunctive antecedents likewise threaten to be a problem for non-trivial '*more*-__-worlds' accounts, which only look at the 'front rows' of $(A \vee B)$-worlds, as ordered by whatever. If the ordering is non-trivial, then for some $A$ and $B$ there will be $A$-worlds earlier in the ordering than any $B$-worlds, and then the $B$-worlds will be ignored—$C$'s truth value at them won't matter. But according to the conclusion of SDA, it does matter; it clearly does in the 'Italy' case. Again, 'most' or 'more' are doing all the work—in this case, all the harm.

But aren't there counterexamples to SDA? What about McKay & van Inwagen's (1977) putative counterexample? Consider the premise:

> "If Spain had joined the Axis or Allies, it would have joined the Axis".

Clearly we don't want to infer from this:

> (Allies $\mathbin{\Box\!\!\rightarrow}$ Axis) "If Spain had joined the Allies, it would have joined the Axis" #

A simple caveat to SDA, which is independently motivated, avoids this problem. Intuitively, a counterfactual presupposes that its antecedent was a live possibility at the relevant time. (Later we will formalise this with its chance being non-zero at an appropriate time; it is analogous to what Bennett (2003, p.55) calls the "zero-intolerance" of indicative conditionals.) So let us restrict SDA to cases in which all the counterfactuals meet this presupposition. Suppose the premise is true. It rules out Spain's joining the Allies as a live possibility (since it claims that Spain *would* have joined the Axis). Hence the presupposition fails for Allies $\mathbin{\Box\!\!\rightarrow}$ Axis; hence this is not a counterexample to SDA, so restricted. We can restrict our attention to problematic cases for 'most __ worlds' accounts for which SDA holds. See also Fine (2012) and Starr (2022) for other rebuttals of such alleged counterexamples to SDA.

In any case, we don't need even my restricted SDA to be universally valid to cause trouble for 'most-__-worlds' accounts. Indeed, we don't need *any* instances of it to be valid. For example, perhaps one has a pragmatic account of why SDA *appears* to be valid, when in fact it is not. (See e.g. Loewer 1976, Schwarz forthcoming.) All that matters is the falsehood of a counterfactual like the 'Rome or Baghdad' one, just as my informants and I intuit, for whatever reason. Appealing to

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

SDA, or its restricted version, helped explain why it comes out false. But never mind SDA; focus on (what I take to be) the counterfactual's falsehood. Any 'most __ worlds' account that judges it to be true is falsified.

We need not assume anything about the (non-trivial) '__' ordering to create this sort of trouble. Here is a general recipe for counterexamples, whatever '__' is: Tell us its ordering of worlds, and in particular which worlds are *most* __. I will then tailor a disjunctive-antecedent counterfactual accordingly that is false, for whatever reason (SDA, restricted SDA, or otherwise). Some disjuncts are verified by the *most* __-worlds, but some are not; the consequent is true at the former but not the latter worlds. The 'most __ worlds' account judges the counterfactual to be true when in fact it is false.

Now let '__' be a *similarity* ordering. Given the generality of the recipe, it too is threatened; all that matters is that it is an ordering with maximally similar $A$-worlds, for any $A$. Perhaps it is an intuitive similarity ordering, one that the folk would readily recognize; perhaps it is more recherché, one that a philosopher proposes for some purpose or other.

## 4.3 | Unspecific antecedents

SDA, perhaps suitably restricted, is a powerful counterexample generator to 'most __ worlds' accounts because it shows how demanding a disjunctive-antecedent counterfactual is: its consequent must be counterfactually true under the supposition of *each* disjunct. 'Most __ worlds' accounts lower the bar for truth: they require only that the consequent be counterfactually true under the supposition of the *most* __ disjuncts. It is thus all too easy for such a counterfactual to clear the bar.

You might think that this turns on a quirk about the word 'or', which does not teach us anything about counterfactuals. But we may get the same conjunctive effect without using the word 'or':

> "If we were in a great ancient city located on a river beginning with 'Ti', we would be in Italy".

There are two such cities: Rome (on the Tiber), and Baghdad (on the Tigris). So this is an 'or'-free way of saying "If we were in Rome or Baghdad, we would be in Italy".[14] And it still sounds false, as most informants that I have asked agree. We see the problem of implausible specificity rearing its head again: the consequent is too specific, given the unspecific antecedent. But again, all I need is the falsehood of the counterfactual (as my informants intuit), for whatever reason.

I have been trying to temper the enthusiasm that one might have for any 'most __ worlds' account, and in particular 'most similar worlds' accounts, based on their verdicts of validity and invalidity of various argument forms. I have argued that we should not agree with their verdict that SDA (suitably restricted) is invalid. But suppose that I am wrong (as are Nute, Fine, Ellis et al, …). Suppose we grant that this verdict is correct, just as 'most similar worlds' accounts say—again, 'most' plays the crucial role in securing this verdict. More power to such accounts, you may say—they are getting even more (in)validity verdicts right than my initial list! There is no doubt that similarity semantics provides an elegant model theory, with soundness and completeness results. (How welcome those results are depends on how welcome the verdicts on validities and invalidities are; I have questioned the SDA verdict but I am granting it now.)

[14] A number of authors have noted that indefinites behave like disjunctions in this regard—see e.g. Alonso-Ovalle (2004).

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

However, since Stalnaker and Lewis are offering *truth* conditions for counterfactuals, it is also crucial that they get right what they take to be the *truth values* of counterfactuals. We have already seen that 'most __ worlds' accounts incorrectly judge various implausibly specific counterfactuals to be true. So 'most' is both friend and foe to 'similarity' accounts—friend to the logic (I am supposing), foe to the truth values.

This brings me to another central concern about similarity accounts of counterfactuals.

## 5 | SIMILARITY

### 5.1 | Intuitive similarity

In the early days of similarity accounts, 'similarity' was regarded intuitively, in accordance with folk judgments. Stalnaker (1980, p.96) speaks of "the extent that an intuitive notion of similarity among possible worlds plays a role" in determining the orderings of worlds. Lewis (1973c, p.420) describes the most similar antecedent-world thus: "an antecedent-world that does not differ gratuitously from ours; one that differs only as much as it must to permit the antecedent to hold; one that is closer to our world in similarity, all things considered, than any other antecedent world". This sounds like a familiar notion of "similarity". On this way of speaking, similarity seems to be a matter of the sharing of properties (perhaps especially natural ones), both locally and globally, and of minimizing discrepancies between the values of salient quantities, among other things. (Recall the "over 7 feet tall" example.) Lewis's earliest works gave us some examples, but not much of a characterization of 'similarity', so it was understandable that it was understood intuitively.

The notion of 'similarity' that allegedly underwrites induction is an intuitive one, albeit difficult to analyze, as Nelson Goodman (1955) made vivid. Indeed, he was scathing of *any* philosophical appeal to 'similarity'. His (1970) begins: "Similarity, I submit, is insidious… Similarity, ever ready to solve philosophical problems and overcome obstacles, is a pretender, an impostor, a quack." What about the notion of 'similarity' that allegedly underwrites counterfactuals? Even nowadays philosophical discussions of counterfactuals often presuppose an intuitive notion of similarity. So *our* discussion of similarity accounts should start with this notion. We will quickly see that it is a non-starter.

Before looking at counterexamples, let me express general puzzlement about this starting point. I think that a typical counterfactual says something about the future relative to a hypothetical scenario—taking some moment in our history, supposing some modification of it specified by the antecedent, and then making a claim about how things evolve from that point onwards. As such, I say that typical counterfactuals are *hypothetical predictions*. "If the cup had been released, it would have fallen" casts us back to a time when the cup's being released was a live possibility; supposing that possibility to be realized, the counterfactual then prognosticates a subsequent event, the cup's falling. ("Would" is the past tense of "will".) Why should (intuitive) *similarity* to the actual world be relevant to this prognostication, let alone the cornerstone?

To be sure, judgments of similarity play an important role in induction—though the recalcitrance of the 'old' and 'new' problems of induction show that even here the notion is somewhat problematic. But nobody builds facts about similarity into the *truth* conditions for claims about the future—predictions about the actual world. Similarity is obviously a non-starter there. That would certainly provide a novel and bold solution to Hume's problem of how one is justified in believing that the future will be similar to the past! Still less does anybody build *maximal* similarity into their truth conditions. Yet according to similarity accounts of counterfactuals, one *can*

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

justify believing that things would happen in a maximally similar way to the actual world, for any counterfactual supposition. After all, their doing so is an analytic truth on such accounts! Much the same points carry over for truth conditions about claims about the past and for backward-looking counterfactuals (e.g. "If a big meteor had struck today, it would have to have been visible yesterday"). I think of such counterfactuals as hypothetical *retrodictions*. Similarity plays no role in claims about the past, so why should it play a role in hypothetical retrodictions?

So when we switch time, similarity is entirely beside the point to the truth conditions. Why should it be, then, that when we (supposedly) switch world, it is regarded as the entire point? Lewis liked to analogize time and modality, and thus tensed discourse and modal discourse. But his injection of similarity in his account of counterfactual discourse strains such analogizing, since it has no such place in any account of tensed discourse.

In any case, there are counterexamples to appealing to intuitive similarity in an account of counterfactuals. Suppose that you have an appointment, and after a long journey you arrive a minute early. Offhand, according to an intuitive similarity relation for counterfactuals, this should be true: 'if you had got lost, you would have arrived on time'. After all, the most *similar* way (to actuality) for you to have got lost would have been for you somehow to have got *minimally* lost—slightly, briefly—and otherwise to have traveled very much as you did. But we can easily fill in the story so that the counterfactual is *not* intuitively true. It could well be that most ways of your getting lost would have caused you to arrive late. Perhaps even the most probable ways of your getting lost would have done so—e.g., at one point of your journey there was an easily-missed turnoff, and missing it leads to a considerable delay. Attending only to those ways that differ the least from actuality ignores the most probable ways—they are too close for comfort. We have seen the tension that may arise between probability and similarity considerations, and we will see it again.

Fine (1975, p.452) has an apparently fatal objection to the intuitive understanding of the similarity relation that underwrites counterfactuals:

> The counterfactual 'If Nixon had pressed the button there would have been a nuclear holocaust' is true or can be imagined to be so. Now suppose that there never will be a nuclear holocaust. Then that counterfactual is, on Lewis' analysis, very likely false. For given any world in which antecedent and consequent are both true it will be easy to imagine a closer world in which the antecedent is true but the consequent false. For we need only imagine a change that prevents the holocaust but that does not require such a great divergence from reality.

Fine is surely right about the ordering given by an intuitive similarity relation. A world in which Nixon's button-pressing is followed by a holocaust is intuitively very *dissimilar* from actuality; much more similar is a world in which the mechanism fails, and thereafter it's business as usual. The recipe for a counterexample, displayed by both my case and Fine's, is as follows. Take a counterfactual in which the consequent is *improbable*, given the antecedent: 'if you had got lost, you would have arrived on time'; 'if Nixon had pressed the button, it would have been business as usual'. But on the intuitively most similar realisation(s) of the antecedent, the consequent is true. The counterfactual seems to be false, but intuitive-similarity semantics judges it to be true.

We may generalize the recipe to a heuristic for searching for counterexamples to similarity semantics: simply delete "intuitively" and "intuitive" from my recipe. Worse still, we may be able to find cases in which the consequent has *minimal* probability, given the antecedent, but the counterfactual comes out true according to the putative similarity relation. We will see this heuristic in action on another putative similarity relation, in §5.3. I do not claim that the heuristic is

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

guaranteed to work (that's why I call it a "heuristic"). For example, a similarity account that is *based on chances* may not succumb to it—see, e.g. Albert (2000, 2014). But again, we are then well on the way to analyzing counterfactuals in terms of chances.

## 5.2 | 'Similarity' as an unspecified ordering

The requisite notion of 'similarity' is not to be found in our scientific or social-scientific theories. Depending on one's goal, one might be happy to regard 'similarity' as a 'black box', an abstract placeholder. One might say that it induces an unspecified ordering of worlds or a selection function, perhaps adding that it is highly context-dependent and vague—and stop there. This may suffice if one wants just to provide a semantics that delivers the pattern of entailments that one wants. Then all that matters is the ordering structure of 'similarity'. In that case I would prefer a word less suggestive of *resemblance*, such as 'closeness', as Stalnaker originally called it, and neo-Stalnakerians like Mandelkern (2019) follow suit. We could just as well call it 'schmimilarity'—and stop there.

But this will not suffice if one wants a theory that delivers intuitively correct truth values to particular counterfactuals. It was crucial for Adams (1975) that we intuit "if Oswald had not killed Kennedy, someone else would have" to be *false*, crucial for Fine that we intuit "if Nixon had pressed the button, there would have been a holocaust" to be *true*, and so on. The latter data point was so important to Lewis that his (1979) system of priorities was largely motivated by it—see the next section. 'Similarity' as a black box makes no predictions about the truth values of particular counterfactuals, and so it gains no confirmation from agreeing with our judgments of those truth values. It also does not help us go *beyond* our judgments, adjudicating hard cases—there are no 'spoils to the victor', because there are no spoils.

To be sure, a semanticist may reply that it is not their job to say which sentences are *true*; rather, it is to say what a sentence's *truth conditions* are. Compare the standard theory of conjunction: '*A* and *B*' is true iff *A* is true and *B* is true. This does not tell us which conjunctions are true; it only tells us *what it takes* for a conjunction *to be true*.[15] However, we have a good independent grip on the right-hand side of this theory—we can regard it as an informative reduction. We do not have such a grip on an unspecified black-box ordering. Those of us with reductionist ambitions should want more. And similarity as a black box will not suffice if one wants our understanding of counterfactuals to further our understanding of other things in which they are thought to be implicated—causation, knowledge, personal identity, perception, laws of nature, dispositions, and so on. For example, Lewis's system of priorities was supposed to undergird the "standard resolution" of the "vagueness" (1979, p.457) of counterfactuals that feature in his analysis of causation.

The less the similarity relation is like intuitive similarity, the more we need to be told what it *is* like, in order to understand which relation it is for these various purposes.

## 5.3 | Lewis's (1979) priorities

And later (1979) Lewis does indeed tell us more. He is motivated by Fine's devastating 'Nixon' counterexample to intuitive similarity being fit for purpose. Lewis replies that this is not the

[15] Thanks here to Matthew Mandelkern.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

relation that is operative in the 'Nixon' counterfactual, or more broadly. He goes on to give his famous and highly influential system of priorities of what matters to the relation:

1. It is of the first importance to avoid big, widespread, diverse violations of law.
2. It is of the second importance to maximize the spatiotemporal region throughout which perfect match of particular fact prevails.
3. It is of the third importance to avoid even small, localized, simple violations of law.
4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly. (472)

This system has faced serious objections—e.g. Elga (2001), Kment (2006). One may wonder: why this seemingly idiosyncratic system of priorities? The priorities seem strangely disunified, crying out for some simpler, unifying principle. We do not find them in other philosophical uses of similarity—for example, in safety and sensitivity in knowledge attributions, or verisimilitude in philosophy of science. And the similarity account of counterfactuals has become more complicated than it was when 'similarity' was simply understood intuitively.

I submit that priority (2) is untenable. In fact, it is not even the conclusion that Lewis himself draws earlier in the same article: "under the similarity relation we seek, *a lot of* perfect match of particular fact is worth a small miracle" (469, my emphasis). Note well: "a lot of" perfect match suggests satisficing rather than maximizing it. The move to *maximizing* perfect match makes matters worse, but in some cases there is not even an inclination to *increase* perfect match with respect to the matter picked out by the antecedent.

Suppose by "last year" I mean the 365-day period that ended at midnight last New Year's Eve, *including* midnight (in my time zone). In fact, I did not click my fingers last year. Consider: 'if I had clicked my fingers some time or other last year, I would have done so *at midnight on New Year's Eve*'.[16] This strikes me as implausibly specific—spectacularly so—but it comes out true by Lewis's lights. After all, the later in the year the hypothetical click takes place, the more perfect match with particular fact is secured, and it is maximized by delaying it to the last possible moment. Let's suppose that you did not click your fingers last year either; neither did Barack Obama nor Madonna; neither did billions of other non-clickers. What would have happened if we had all clicked our fingers some time or other last year? By the lights of Lewis's priorities, we *all* would have done so at the stroke of midnight (in my time zone), in perfect sync. Quite a way to ring in the New Year! But never mind postponing all the clicks to the last possible moment. There is no pressure to delay the hypothetical clicks *at all*. For all we care, and for all world-similarity should care, they could have happened at any times in the year—such is the unspecificity of the antecedent. Recall again how Lewis objected to Stalnaker's account favoring one specific nationality over another in 'If Bizet and Verdi were compatriots …' We see that Lewis's account also betrays favoritism—indeed, far more strikingly here.

Now suppose that "last year" does *not* include the final midnight (the usual meaning, I think)—it is a time interval that is open on the right. Then we cannot maximize the spatiotemporal region

[16] Here I strengthen an example due to Pollock (Nute, 1982): I forgot my coat in a bar last night. Two potential thieves passed by, one at 10 pm, another at midnight, each with a positive chance of stealing the coat, but they did not. The counterfactual 'if my coat had been stolen last night, it would have been stolen at midnight' comes out true by Lewis's lights. My example is simpler but even more absurd. Rather than just two potential times at which the hypothetical event could have taken place, there are *uncountably many*; yet the counterfactual privileges exactly one of them. What follows develops the point further.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

throughout which perfect match of particular fact prevails. Instead, there are larger and larger regions that delay my click time (and yours, etc.) more and more: until the final midnight – 1/n seconds, for n = 1, 2, 3, … Priority (2) cannot be met. (This is reminiscent of Lewis's "more than 7 feet tall" example in which the limit assumption is not met.) In such cases it should not be a priority at all.

Lewis proposes his (1979) priority system for deterministic worlds. He assumes that the base world from which the Nixon counterfactual is evaluated is *deterministic*. It is a world that "may or may not be ours" (467). This is a little surprising, because Fine's example is clearly to be evaluated from *our* world, and he makes no assumption that our world is deterministic. On the contrary, we have some reason to say that Lewis's base world is *not* ours, since we have some reason to think that ours is indeterministic, and he elsewhere affirms this explicitly (e.g., 1980). To be sure, it is a major contribution to specify what matters to similarity for deterministic worlds. And Lewis (1986b) considers indeterministic worlds in a postscript to the paper, several years later—more on that shortly.

Indeterminism introduces further problems. Suppose again that radioactive decay is indeterministic (as Lewis himself does). Here is a tritium atom that has not decayed. Consider: 'If the atom had decayed last year (the final midnight included), it would have done so at midnight on New Year's Eve.' The event that would have happened by Lewis's lights had chance 0. We can quantify how implausibly specific it is: maximally so! As before, there should be no favoritism *at all* towards later decay times. In fact, if anything there should be favoritism towards *earlier* decay times. After all, the atom's decay time follows an exponential probability density function, which peaks at the *start* of last year and decreases monotonically thereafter, hitting its *minimum* at the end of the year.

To be sure, every exact decay time has probability zero[17]—so a counterfactual that singles out any particular time is implausibly specific (an understatement!). But now suppose we subdivide last year into intervals of one second. All the intervals have positive probability of containing the hypothetical decay time, but the final one has uniquely the *lowest* probability. And yet according to priority (2), that's the one in which the atom would have decayed. 'If the atom had decayed last year, it would have decayed in the final interval', comes out true according to Lewis's similarity relation, but the consequent has *minimal* probability, given the antecedent. This showcases the heuristic for generating counterexamples to similarity accounts that I offered earlier. Similarity and probability are pulling in opposite directions—maximally so. Probability and counterfactuals are closely related, but not like *this*! At the end I will offer an account in which they are in harmony.

## 5.4 | Indeterminism: quasi-miracles

A fully general theory of counterfactuals needs to apply under indeterminism also. This is especially pressing, since we have some reason to think that the actual world is indeterministic, and Lewis himself believes this. And he does worry about the system of priorities under indeterminism in (1986b, 58–65). He is concerned that if our world is indeterministic, there could be perfect reconvergence to actual history after Nixon presses the button without any miracles: "if chance processes are abundant, … why couldn't they accomplish the cover-up?". So he refines his priorities, introducing the notion of a *quasi-miracle*:

[17] Assuming real-valued probabilities, as the usual formulation of the decay law does.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

> a quasi-miracle… though it is entirely lawful, nevertheless detracts from similarity… The quasi-miracle would be such a remarkable coincidence that it would be quite unlike the goings-on we take to be typical of our world… the chance outcomes seem to conspire to produce a pattern (60).

Cover-up, conspiracy—how appropriate in an example about Nixon! (But he was not entirely lawful.) More generally, quasi-miracles are meant to help deliver the intuitive truth values of counterfactuals that would otherwise be undermined by freakish patterns of goings-on.

Talk of chance outcomes that "seem to conspire" is obviously metaphorical. "Remarkable" is not a scientific notion to the extent that it goes beyond "low chance", and it's important for Lewis that it does.[18] It sounds anthropocentric, a matter of what *we* take to be (a)typical. Yet Lewis (1986b, p.59) disparages quantum mechanics' "unfortunate anthropocentric foundation" (referring to the reduction of the wave function due to measurement). Depending on who "we" are, penalising "remarkable" outcomes for similarity may even be anti-scientific: quantum mechanics and relativity are remarkable to folk thinking, and in EPR correlations between space-like separated particles, the particles seem to conspire. And the notion of a "coincidence" also has anthropocentric associations, as when Aristotle describes a meeting of friends at a market that one regards as lucky, even though the presence of each of them has a complete causal explanation. Lewis was at pains to give counterfactuals a secure foundation so that they could underpin various objective notions—recall his counterfactual analyses of causation, the direction of time, and personal identity. It is unclear to me how a notion of 'similarity' that appeals to quasi-miracles can bear this weight.

Lewis's similarity account has become more complicated: quasi-miracles have been introduced to handle counterfactuals under indeterminism. A quasi-miracle detracts from similarity, but how much? How do we compare a big, widespread, diverse quasi-miracle with a medium-sized genuine miracle? How does perfect match of particular fact trade off against a large quasi-miracle? Questions such as these are left unanswered—channel the spirit of Goodman. And to the extent that one could add further details to answer them, the original analysis would become still more complicated.

Lewis's appeal to quasi-miracles is uncomfortably reminiscent of *Cournot's Principle*, a principle with a long history in the philosophy of probability literature. One version of it says that no improbable event will happen. I find this principle to be implausible, wherever we set the bar for 'improbable'. Whatever a given tritium atom ends up doing will have probability 0 according to its exponential decay law—either its decaying at an exact moment, or its never decaying. I also find implausible a similar principle about quasi-miracles, which I dub:

> *Quasi-Cournot's Principle*: No quasi-miracle *will* happen.

Consider a long sequence of independent trials of tossing a fair coin. Suppose that 19 heads in a row do not constitute a quasi-miracle, but that 20 heads in a row do. (We may easily change those numbers if you like, and we can allow that the cut-off is vague.) Now suppose that the coin happens to land 19 heads in a row; then by Quasi-Cournot's Principle, the 20$^{th}$ heads *will not*

[18] He writes: "My point is not that quasi-miracles detract from similarity because they are so very improbable. They are; but ever so many unremarkable things that actually happen, and ever so many other things that might happen under various counterfactual suppositions, are likewise very improbable. What makes a quasi-miracle is not improbability per se, but rather the remarkable way in which the chance outcomes seem to conspire to produce a pattern" (60).

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

*happen*. This is already absurd: it is to deny that the 20th toss is independent of the tosses that came beforehand. It is tantamount to attributing a memory to the coin.

Quasi-Cournot's Principle also licenses the gambler's fallacy: if you witness a sufficiently long sequence of heads, it becomes rational for you to bet on *tails* on the next toss. After all, according to the Principle, heads *will not happen*, as its doing so would complete a quasi-miraculous event.

Quasi-Cournot's Principle is no better when we make it counterfactual:

> No quasi-miracle *would happen* (if *A* were the case—insert your favourite non-quasi-miraculous antecedent *A*).

For example, here is a coin that is never tossed; but suppose it were tossed 20 times. 'If it were to land heads on the first 19 tosses, it *would not land heads on the* 20th' is implausible—yet again, implausibly specific. And the reasoning behind the gambler's fallacy is no less fallacious when we go counterfactual. But the quasi-miracle approach is committed to both. Let me emphasize again my view of counterfactuals as hypothetical predictions: Quasi-Cournot's Principle and the gambler's fallacy are specious for predictions, and so they are for hypothetical predictions.

## 5.5 | Williams: Typicality

Lewis says that quasi-miracles are "quite unlike the goings-on we take to be *typical* of our world". Williams (2008) develops this idea. He replaces the notion of quasi-miracles with that of *atypicality*. Each possible sequence of fair coin tosses is equally probable, but most sequences are *typical*, while a sufficiently long sequence of heads (say) is *atypical*. "Seems to conspire to produce a pattern" sounds like *non-random*. Williams proposes: "[W]e can identify the required notion of typicality (relative to an assignment of chances) with the mathematical property of a set of outcomes being random" (407-408). Randomness, in turn, can be given a rigorous and objective characterisation. My concerns about unclarity and anthropocentrism are dispelled—the notion has good scientific credentials. *Atypical* (non-random) patterns detract from similarity: worlds in which they occur are less similar than worlds in which they do not.

However, I have concerns about the appeal to typicality in the truth conditions for counterfactuals. Consider another questionable Cournot-like principle:

> *No atypical event will happen*.

This is surely false: atypical events have happened in the past, and they surely will do so in the future. (All the more so if the future is longer than the past, as most cosmologists believe.) A counterfactual version of the principle does not seem any more plausible to me:

> *For any antecedent A, if A were the case, no atypical event would happen*.

I don't think that is true—in fact, it would be very atypical for there to be no atypical events! I especially don't think that it is *analytically* true. Yet this seems to be a commitment of Williams' view.[19]

[19] The argument that follows is reminiscent of one in Hawthorne (2005, p.402) against Lewis's appeal to quasi-miracles, which also appeals to agglomeration of many specific counterfactuals.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Assume the universe is finite, and partition it into a huge finite number $n$ of small regions of space-time, $R_1$, $R_2$, …, $R_n$. Let $A$ specify some typical non-actual event—say, my blinking at a particular time. What would happen if $A$ were the case? According to Williams' account, nothing atypical would happen in $R_1$. After all, $A$-worlds in which something atypical happens in $R_1$ are less similar than typical worlds. Likewise for $R_2$ … Likewise for $R_n$. By agglomeration, we have:

> *If A were the case, nothing atypical would happen in* $R_1$ *and nothing atypical would happen in* $R_2$ *and … and nothing atypical would happen in* $R_n$.

Replacing the consequent with a simpler equivalent, and remembering that the $R_i$ regions collectively fill the entire universe, we have

> *If A were the case, nothing atypical would happen anywhere in the universe.*

This is an instance of the counterfactual version of the questionable Cournot-like principle.

Now assume that the universe is infinite. Then we may partition it with a huge countable partition of small regions of space-time and appeal to countable agglomeration. Or better still, simply take some enormous finite part of the universe, form a huge finite partition of *that*, and run the previous argument.

And as before, the gambler's fallacy is counterfactually vindicated by the (a)typicality account—and it shouldn't be. Or perhaps we could replace typicality by *normality* in the account of similarity. (Normality is enjoying considerable attention in epistemology and philosophy of science—see e.g. Goldstein & Hawthorne, 2022; Goodman & Salow 2023). I would need to see the details, but I wager that I could rewrite these arguments against such an account.

A number of variants of similarity accounts have been offered—e.g. by Bennett (2003), Schaffer (2004), Kment (2006), and Nolan (2017). Some specific criticisms that I have given may or may not apply; I do not have the space to consider each variant. But I have raised deeper problems that I believe generalize to any account structurally like Stalnaker's and Lewis's. This casts doubt on the entire approach. If similarity accounts were correct, the facts about counterfactuals would be very different from what I take them to be.

## 6 | CONDITIONAL CHANCES INSTEAD OF SIMILARITY

This completes my critique of most-similar-worlds accounts of counterfactuals, and I could end this essay here. However, I want to conclude more positively, briefly championing the direction in which I believe the analysis of counterfactuals should go. Chances have repeatedly been working their way into our discussion, given their close connection to counterfactuals. It is time to bring them to center stage, and to recognize this connection explicitly.

Early on we saw the famous Ramsey quote that understands conditionals in terms of conditional probabilities. Conditional credence accounts of indicative conditionals have been highly popular—especially Adams' Thesis regarding their probability or assertability (Adams, 1975; Lewis, 1976). I voiced the strong presumption that our accounts of indicative conditionals and counterfactuals should be as unified as possible. As I said, the ontological analogue of a conditional credence is a *conditional chance*. Moreover, conditional chances fit neatly with my view of counterfactuals as typically making hypothetical predictions. And so it is natural to understand counterfactuals in terms of conditional chances instead of most-similar worlds.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Skyrms (1984), Edgington (2008), Leitgeb (2012a, 2012b), Kvart (2015), Loewer (2020), and Fernandes (MS) also do so. There are some intramural disputes between us, but they are not important here, and they are minor compared to our differences with similarity accounts. The rough idea that we share is that a counterfactual that we rightly assert or affirm corresponds to a high conditional chance of its consequent given its antecedent. For example, Leitgeb offers this truth condition, where *ch* is the chance function:

> $A \mathbin{\square\!\!\rightarrow} C$ iff $ch(C|A)$ is very high.

He regards "very high" as context-sensitive. The chance function is implicitly time-indexed. I set the threshold as high as possible, so it is not context-sensitive, and I make the time reference explicit:

> $A \mathbin{\square\!\!\rightarrow} C$ iff $ch_t(C|A) = 1$ at a time $t$ shortly, but not too shortly, before the time or period picked out by $A$.

If a conditional chance ever becomes 1, it stays 1 thereafter.

This provides a truth condition for counterfactuals evaluated in the actual world. More generally, we can index both sides of my biconditional to any base world *w*, much as Stalnaker and Lewis did. We simply take the chance function to be implicitly world-indexed also.

More pithily:

> $A \mathbin{\square\!\!\rightarrow} C$ iff $ch_t(C|A) = 1$ at a fork time $t$ associated with $A$.

This terminology is inspired by Bennett (2003, p.216). Likewise, Edgington (2004, p.13) speaks of "a point of deviation" or a "fork" that is "shortly before the antecedent-time" in the Lewis-style truth conditions for counterfactuals that she envisages (but does not endorse). But I make no appeal to similarity in my reference to "a fork time $t$ associated with $A$".[20] It is just shorthand for the wordier expression that it replaces, evocative of a time when things could have gone either way with respect to $A$. (If the conditional chance is 1 at any candidate fork time, it remains so at all subsequent candidates.) And my referring to a "time or period" rather than only a single instant, "the antecedent-time", also handles antecedents that span intervals of time, such as "the atom decays last year".

My time reference is also like Edgington's in her own chance-based account:

> "If A turns out to be false, the objectively correct value to be assigned to the counterfactual, 'If it had been the case that A, it would have been the case that C' is the conditional chance of C on the supposition that A … *just before it turned out* that A" (p.16, my italics).

And my time reference is reminiscent of the classic Skyrms' Thesis about the assertability or probability of a counterfactual. Santorio (forthcoming, p.6) puts it thus:

[20] I pass over some technical points regarding probability 1 being only 'almost sure' rather than 'sure' for real-valued probabilities, à la Kolmogorov. I could fix this problem by appealing to comparative conditional probability (Koopman 1940) and making the conditional probability "maximal" rather than "1", or by appealing to infinitesimal probabilities (see Wenmackers 2019).

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

> Skyrms' Thesis involves a shifted time-index on the chance function: we should consider not the current chances, but rather the chances that obtained at some point in the past. In particular, we should consider the chances that obtained 'just before' the truth status of the counterfactual antecedent was settled. This introduces an element of vagueness in Skyrms' Thesis—exactly what time should we pick out?—but this vagueness is generally taken to be tolerable.

However, I do not assume that the antecedent turned out to be false, or that it was settled.

We are all addressing the problem of allowing what Bennett (2003, p.214) calls a "*ramp* from the actual world to the antecedent of the conditional", which starts at the fork time. We want to keep much of actual history fixed, avoiding gratuitous changes to the distant past—hence, our locutions "shortly before" or "just before". But it is important also to add my rider "but not too shortly": we also want to avoid abrupt, last-moment transitions that bring about the antecedent. (Think of a well-constructed ramp on to a freeway that does not start too early, but also not too late, as is the case for the white-knuckle entries onto the Pasadena Freeway, Interstate 110.) In Fine's 'Nixon' counterfactual, for example, we don't imagine Nixon suddenly lunging for the button at the last second, but rather approaching it more smoothly.

I can allow some vagueness (as Santorio says) and context-sensitivity concerning the fork time, which gives it some flexibility. But I think that as a useful heuristic we typically imagine it to be a time when the chance of the antecedent was maximal (or close to it). Typically, when contemplating a counterfactual we imagine giving its antecedent its best shot at being realized. For example, we imagine a time when Nixon was most likely to push the button—say, when the Cold War was at its height during his presidency. If there are multiple such times for a given counterfactual, then the corresponding conditional chance needs to be 1 at all such times. (If it was less than 1 at some of them, then the counterfactual is hostage to its chanciness being resolved the right way.) However, I have not built this heuristic into my truth condition, keeping it more flexible. In any case, all theories of counterfactuals face the ramp problem, with better or worse solutions to it.

The requirement that $ch(C \mid A) = 1$ ensures the validity of various argument forms. For example, it is trivial to show that my account delivers the validity of Agglomeration. Suppose that Agglomeration's premises hold, with the chance function at $A$'s fork time $t$:

$$ch_t\,(C_1|A) = 1$$

$$ch_t\,(C_2|A) = 1.$$

It follows immediately that

$$ch_t\,(C_1\&C_2|A) = 1,$$

at $t$, my analysis of Agglomeration's conclusion. Indeed, I may uphold the validity of countable Agglomeration. My account trivially upholds the validity of modus ponens (at least for conditional-free antecedents and consequents). It also upholds the validity of SDA restricted to

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

counterfactuals whose antecedents were live possibilities (since the chance of such antecedents is positive).[21]

Furthermore, my account delivers the *invalidity* of antecedent strengthening. This is due to a shift in the operative chance function. A strengthened antecedent often requires a different fork time: the fork time associated with $A$ & $B$ may be different from that associated with $A$. And my account straightforwardly delivers the invalidity of transitivity and contraposition, agreeing with Stalnaker's and Lewis's accounts there.

We may also define a probabilistic counterfactual that formalizes 'if $A$ were the case, then it's $p$ probable that $C$ would be the case':

$$A \mathrel{\square\!\!\rightarrow_p} C \text{ iff } ch_t(C|A) = p \text{ at a fork time } t \text{ associated with } A.[22]$$

For example:

If a fair coin had been tossed, it's 50% probable that it would have landed heads.

English expressions of the form 'if $A$ were the case, then the chance of $C$ would be $p$' and 'the chance that if $A$ were the case, then $C$ would be the case, is $p$' have a reading on which they express $A \mathrel{\square\!\!\rightarrow_p} C$.

I find it plausible that plain 'would' should be formalized as 'would certainly'. If $C$ were chancy if $A$, then 'if $A$ were the case, $C$ would be the case' is false. This plain 'would' counterfactual overreaches, second-guessing the chancy resolution of $C$. Nothing less than certainty of the consequent will do for the truth of the 'would' counterfactual. Chance undermines 'would'! Then this definition generalizes my truth condition, with plain 'would' counterfactuals being the special case in which $p = 1$. It also gives the right results for counterfactuals like 'if the tritium atom had decayed last year, it would have been more likely to do so earlier in the year than later in the year', with higher values of $p$ for earlier time intervals of the same length.

Chance-based accounts of counterfactuals are explanatory: they tell us *why* counterfactuals are true. Counterfactuals are grounded in facts about chances. We need such facts anyway—indeed, physicists work hard to figure out some of them. The accounts are also simple—much more so than Lewis's after his priorities for what matters for similarity under determinism and indeterminism are spelled out. Moreover, our accounts need not go silent where his does. And they provide a reduction, identifying what feature of the actual world determines a counterfactual's truth value: the value of the corresponding conditional chance. Chances are world-based according to the leading accounts of chance, and it is the chance in the actual world that matters to the truth of the counterfactuals that we utter. According to actual frequentism, chances are actual relative frequencies—for example, the chance of heads is ½ iff half of the actual coin tosses (in some suitable reference class) land heads. Likewise, best-system analyses, which are sophistications of frequentism, regard chances as the probabilities that appear in the theory of the actual world that best balances simplicity, strength, and fit. Propensity interpretations

[21] I take the fork time associated with $A \vee B$ to be the earlier of the fork times associated with $A$ and with $B$ if those times differ (otherwise, they are all the same time). Suppose that $ch(C \mid A \vee B) = 1$ at that time. Then $ch(C \mid A) = 1$ and $ch(C \mid B) = 1$ at that time, and they remain 1 thereafter.

If one maintains that SDA (even restricted) is invalid, one could still uphold the spirit of my chance account with some suitable tweak to its time reference: its pithier statement can still serve as a template, allowing variations in the putative fork time. One might even leave it as a 'black box' if one so desires. Recall also that for counterfactuals like 'if we were in Rome or Baghdad, we would be in Italy' to be problematic for similarity accounts, I did not need (restricted) SDA to be valid; I just needed the counterfactuals to be false, for whatever reason.

[22] I am grateful here to Wolfgang Schwarz.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

portray chances as measuring graded dispositions or tendencies of parts of the actual world. According to the method of arbitrary functions, chances capture certain symmetries in the actual evolution functions of dynamic systems. (See Hájek, 2023 for far more detail on these accounts.)

Recall that "most" and "similar" each caused problems for most-similar worlds accounts. Analogous problems do not arise for my account. What about "most"? Recall the problems of implausible specificity for most–similar–worlds accounts, or indeed any most–__–account. This was due to their favoritism for only the maximal members of an ordering of worlds, ignoring worlds that are 'further back'. There is no parallel problem for my chance-based account. All worlds that have positive chance at the relevant time have their day in the sun; and typically those that have no chance are rightly ignored.[23] And yet my chance account makes the nuanced predictions that it should, as in the tritium atom example.

What about "similar", in the spirit of Stalnaker and Lewis? This is a fundamental point with which chance-based accounts disagree. I have repeatedly pointed out tensions between probability and similarity considerations. My account cleaves to the former rather than the latter, as it should. Non-actual possibilities that have extremely low chance might be very 'similar' (in some good sense)—think of all the possible decay times of a tritium atom that actually decayed at noon. And less 'similar' possibilities might have higher chance—the atom's *not* decaying at noon had chance 1, but it is less similar than what actually happened according to strong centering (each world is more similar to itself than any other world is to it), which both Stalnaker and Lewis assume.

Now perhaps one might be able to define some notion of 'similarity' *in terms of chances*, so that my chance-based account could be recast as a 'similarity' account after all. If after some contortions one can mimic my account with an ordering and then insist on calling it a 'similarity' account, so be it. But it would be an unilluminating and even misleading way to describe it—a strained notion of 'similarity' that suppresses what is distinctive about my account, placing chance at center stage. It would only bring home the concern that so-called 'similarity' may be such a broad, nebulous word that one can apply it promiscuously—it is well on its way to being a black box. It would leave the spirit of the original Stalnaker/Lewis accounts far behind. Far better to cut to the chase and call my account what it is: a chance-based account, with substantive constraints on what that means.

Recall the concern about the unclarity of 'similarity', and Goodman's withering assessment of it. One might worry that 'chance' is similarly unclear. But I submit that it is in far better order than 'similarity'. For starters, probability theory is a rich and highly developed area of mathematics, a handmaiden of the sciences and the social sciences, in a way that similarity certainly is not. Chance is a kind of probability, and its formalism is far better understood than that of similarity.[24] Now, I cannot settle here what is the best understanding *of* chance, but I have the vast literature on the interpretations of objective probability to draw on—I have already begun to draw on it above. To be sure, there are still problems to be solved and details to be filled in; but

[23] I say "typically" because counterfactuals about what would have been the case if the chances had been different may pose a problem—but they are highly atypical. In any case, counterlegals pose a somewhat parallel problem for Lewis's priority system.

[24] Indeed, Jeremy Goodman (MS, no relation!) argues that Lewis's (1973b) official analysis of counterfactuals cannot be understood in terms of comparative similarity, given some of the latter's apparent formal properties—yet of course that is exactly how many authors including Lewis himself understand it.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

let's not hold chance to a higher standard than similarity, which has problems and missing details in spades.

Again, the sciences and social sciences traffic freely in the notion of chance, so it has earned its keep and then some, though they have not settled on an account *of* chance. Chance also plays a major role in various parts of philosophy, especially in philosophy of science, metaphysics, and formal epistemology (which is enjoying its own golden age!). Indeed, we can use the multiple roles to help fix the referent of 'chance', far more determinately than we can fix the referent of 'similarity'. So we should embrace chances anyway, and understanding them better is a worthy project. Once we reduce counterfactuals to chances, we spare ourselves the further project of understanding 'similarity'.

One may object to my account that it leads to counterfactual skepticism—my truth conditions are very demanding. For example, by my lights "if the cup had been released, it would have fallen" is false. The conditional chance of its falling, given its release (at the relevant time) was less than 1—there was some chance of its being lifted by an unexpected updraft, or of someone suddenly reaching over to save it from falling, or what have you. Indeed, it is exactly to avoid counterfactual skepticism that Leitgeb sets the threshold for 'very high' lower than I do. Again, this is not the place to get into that debate in any detail—I gladly do so elsewhere (Hájek, MS). I argue that far from being a cost of my account, counterfactual skepticism is independently plausible. Chance undermines 'would'! Unspecificity undermines 'would'! And most counterfactuals that we utter or affirm involve either antecedents that are unspecific or consequents that are chancy (or both).

Moreover, I have a simple error theory of why we take so many of our cherished counterfactuals to be true, when in fact they are false: they may well be *approximately true*. The corresponding conditional chances may be *approximately* 1. (Notice that unlike similarity, here there is only one dimension of closeness: differences between numbers representing probabilities.) For example, "if the cup had been released, it would have fallen" is approximately true because the corresponding conditional chance is very close to 1. Normal conversation often permits loose talk: we may say things that are strictly speaking false, but close enough to true for our purposes—so close that we may even take them to be true. We may say in a somewhat relaxed context "France is hexagonal", and even regard what we say as true, when strictly speaking it is false. We don't always speak strictly.[25]

Indeed, this showcases yet another benefit of a chance-based account over similarity accounts. Since chances come in degrees, we can easily make sense of *closeness to truth* for counterfactuals; this is a harder task for similarity accounts. In any case, as Leitgeb's account shows, many of the benefits of a chance-based account over a similarity account can be had without leading to skepticism, and I do not want to insist on my account here.

I cannot spell out here the details of a chance-based account of counterfactuals and to defend it against possible objections—that must be postponed to another occasion. My goal in this section has been more modest: to motivate such an account, to sketch its contours, and to argue that it is

[25] It does not follow that *whenever* a counterfactual is approximately true, we will take it to be true. Sometimes a sentence itself raises the standards for what we take to be true—for example, by drawing our attention to unlikely possibilities, or by marking a need for precision. "If we were in Rome or Baghdad, we would be in Italy" is approximately true, I assume: the relevant conditional chance of our being in Italy, given that we are Rome or Baghdad, is presumably very high, given our dispositions and so on. But the very mention of Baghdad forces us not to ignore the (unlikely) possibility of our being there, and we should not intuit that the counterfactual is true. Compare: we should not intuit that "there were *exactly* 100 people at the party" is true when we know that there were 98 people there, even though we know that the claim is approximately true. Sometimes near enough is not good enough, even though it often is. (Thanks to a referee for *Philosophy and Phenomenological Research* for pressing me here.)

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

well-placed to make progress on the numerous problems that beset most-similar-worlds accounts of counterfactuals. My overall goals have been to make the case in the bulk of this paper for a paradigm shift away from such accounts, and then to indicate the direction that the shift should take.

## REFERENCES


Adams, E. (1975). *The logic of conditionals*. Reidel.

Albert, D. Z. (2000). *Time and chance*. Harvard University Press.

Albert, D. Z. (2014). The sharpness of the distinction between past and future. In A. Wilson (Ed.), *Asymmetries of chance and time*. Oxford University Press.

Alonso-Ovalle, L. (2004). Simplification of disjunctive antecedents. In K. Moulton & M. Wolf (Eds.), *Proceedings of the 34th North East Linguistic Society Conference*. University of Massachusetts, GLSA, pp. 1–15.

Bennett, J. (2003). *Conditionals*, Oxford University Press.

Berntson, D. (MS). The paradox of counterfactual tolerance.

Bradley, R. (2012). Multidimensional possible-world semantics for conditionals. *Philosophical Review*, *121*(4), 539–571.

Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies*, *160*(1), 139–166.

Eagle, A. (2011). Deterministic chance. *Noûs*, *45*(2), 269–299.

Edgington, D. (2004). Counterfactuals and the benefit of hindsight. In P. Dowe & P. Noordhof (Eds.), *Chance and cause* (pp. 12–27). Routledge.

Edgington, D. (2008). Counterfactuals. *Proceedings of the Aristotelian Society*, *108*(1), 1–21.

Elga, A. (2001). Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science*, *68* PSA (2000), S313–S324.

Ellis, B., Jackson, F., & Pargetter, R. (1977). An objection to possible-world semantics for counterfactual logics. *Journal of Philosophical Logic*, *6*, 355–357.

Fernandes, A. (MS). The branchpoint proposal and the role of counterfactuals.

Fine, K. (1975). Critical notice: *Counterfactuals*. *Mind*, *84*, 451–458.

Fine, K. (2012). Counterfactuals without possible worlds. *Journal of Philosophy*, *109*(3), 221–246.

Goldstein, S., & Hawthorne, J. (2022). Counterfactual contamination. *Australasian Journal of Philosophy*, *100*(2), 262–278.

Goodman, J., & Salow B. (2023). Epistemology normalized. *Philosophical Review*, *132*(1), 89–145.

Goodman, J. (MS). Counterfactuals and comparative similarity. http://www-bcf.usc.edu/jlgoodma/CounterfactualsWithoutCloseness.pdf

Goodman, N. (1955). *Fact, fiction, and forecast*. Harvard University Press.

Goodman, N. (1970). Seven strictures on similarity. In L. Foster & J. W. Swanson (Eds.), *Experience and theory*. University of Massachusetts Press.

Günther, M., & C. Sisti (2022). Ramsey's conditionals. *Synthese*, *200*(165), 1–31.

Hájek, A. (2021). Contra counterfactism. *Synthese*, *199*, 181–210.

Hájek, A. (2023). Interpretations of probability. In E. N. Zalta & U. Nodelman (Eds.). *The Stanford encyclopedia of philosophy* (Winter 2023 ed.), https://plato.stanford.edu/archives/win2023/entries/probability-interpret/.

Hájek, A. (MS). Most counterfactuals are false.

Harper, W. L., Stalnaker, R., & Pearce, G. (Eds.) (1981). *Ifs*. D. Reidel.

Hawthorne, J. (2005). Chance and counterfactuals. *Philosophy and Phenomenological Research*, *70*(2), 396–405.

Hoefer, C. (2019). *Chance in the world: A Humean guide to objective chance*. Oxford University Press.

Ismael, J. (2011). A modest proposal about chance. *Journal of Philosophy*, *108*(8), 416–442.

Jackson, F. (1987). *Conditionals*. Blackwell.

Khoo, J. (2022). *The meaning of If*. Oxford University Press.

Kment, B. (2006). Counterfactuals and explanation. *Mind*, *115*(458), 261–310.

Koopman, B. O. (1940). The axioms and algebra of intuitive probability. *Annals of Mathematics*, *41*, 269–292.

Kratzer, A. (2012). *Modals and conditionals: New and revised perspectives*. Oxford University Press.

Kroedel, T., & F. Huber (2013). Counterfactual dependence and arrow. *Noûs*, *47*(3), 453–466.


19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Kvart, I. (2015). The causal-process-chance-based analysis of counterfactuals. https://philarchive.org/rec/KVATCA-2
Leitgeb, H. (2012a). A probabilistic semantics for counterfactuals. Part A. *Review of Symbolic Logic*, *5*(1), 16–84.
Leitgeb, H. (2012b). A probabilistic semantics for counterfactuals. Part B. *Review of Symbolic Logic*, *5*(1), 85–121.
Levi, I. (1990). Chance. *Philosophical Topics*, *18*(2), 117–149.
Lewis, D. (1973a). Causation. *Journal of Philosophy*, *70*, 556–567.
Lewis, D. (1973b). *Counterfactuals*. Blackwell and Harvard University Press.
Lewis, D. (1973c). Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, *2*, 418–446.
Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, *85*(3), 297–315.
Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*, 455–476.
Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. II, pp. 263–293). University of California Press.
Lewis, D. (1986a). *On the plurality of worlds*. Blackwell.
Lewis, D. (1986b). *Philosophical papers* (vol. II). Oxford University Press.
Lewis, D. (1994). Humean supervenience debugged. *Mind*, *103*(412), 473–490.
Loewer, B. (1976). Counterfactuals with disjunctive antecedents. *Journal of Philosophy*, *73*, 531–537.
Loewer, B. (2001). Determinism and chance. *Studies in the History of Modern Physics*, *32*(4), 609–620.
Loewer, B. (2020). The mentaculus vision. In V. Allori (Ed.) *Statistical mechanics and scientific explanation: Determinism, indeterminism, and laws of nature*, World Scientific, 3–29.
MacFarlane, J. (2014). *Assessment sensitivity: Relative Truth and Its Applications*, Oxford University Press.
Mandelkern, M. (2019). Talking about worlds. *Philosophical Perspectives*, *32*, Philosophy of Language.
McKay, T., & van Inwagen, P. (1977). Counterfactuals with disjunctive antecedents. *Philosophical Studies*, *31*, 353–356.
Morreau, M. (2010). It simply does not add up: Trouble with overall similarity. *Journal of Philosophy*, *107*(9), 469–490.
Nichols, C. (2021). Relevance first: Relocating similarity in counterfactual semantics, *Synthese*, *198*, 10529–10564.
Nolan, D. (2003). Defending a possible-worlds account of indicative conditionals. *Philosophical Studies*, *116*(3), 215–69.
Nolan, D. (2017). Causal counterfactuals and impossible worlds. In H. Beebee, C. Hitchcock, & H. Price, (Eds.), *Making a difference*. Oxford University Press, 14–32.
Nute, D. (1975). Counterfactuals and the similarity of words. *Journal of Philosophy*, *72*(21), 773–778.
Nute, D. (1982). Topics in conditional logic. *Mind*, *91* (361),136–138.
Nute, D., & Cross, C. B. (2001). Conditional logic. In D. M. Gabbay & F. Guenthner (Eds.), *Handbook of philosophical logic* (vol. 4). Kluwer, 1–98.
Pollock, J. L. (1976). The 'possible worlds' analysis of counterfactuals. *Philosophical Studies*, *29*(6), 469–476.
Popper, K. (1959). The propensity interpretation of probability. *British Journal of the Philosophy of Science*, *10*, 25–42.
Ramsey, F. P. (1928/1990). General propositions and causality. In D. H. Mellor (Ed.), *Philosophical papers*. Cambridge University Press (pp. 145–163).
Santorio, P. (2017). Conditional excluded middle in informational semantics. In A. Cremers, T. van Gessel, & F. Roelofsen (Eds.), *Proceedings for the 21st Amsterdam Colloquium*, 385–394.
Santorio, P. (2018). Alternatives and truthmakers in conditional semantics. *Journal of Philosophy*, *115*(10).
Santorio, P. (2019). Interventions in premise semantics. *Philosopher's Imprint*, *19*(1).
Santorio, P. (forthcoming). Probabilities of counterfactuals are counterfactual probabilities. Forthcoming in *Journal of Philosophy*. Page reference to: https://paolosantorio.net/chancecount.jphil.pdf
Schaffer, J. (2004). Counterfactuals, causal independence and conceptual circularity. *Analysis*, *64*, 299–309.
Schaffer, J. (2007). Deterministic chance? *British Journal for the Philosophy of Science*, *58*(2), 113–40.
Schulz, M. (2017). *Counterfactuals and probability*. Oxford University Press.
Schwarz, W. (MS). Unspecific antecedents. In B. Fitelson & J. J. Joaquin (Eds.), *Could, would, should: Essays in honour of Alan Hájek*. Springer Nature.
Slote, M. (1978). Time in counterfactuals. *The Philosophical Review*, *87*(1), 3–27.
Skyrms, B. (1984). *Pragmatics and empiricism*. Yale University Press.
Stalnaker, R. C. (1968). A theory of conditionals. *Studies in logical theory, American philosophical quarterly, Monograph: 2*, 98–112. Blackwell. Reprinted in Harper et al. (Eds.) (1981), 41–55; page references to this volume.

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Stalnaker, R. C. (1981). *A defense of conditional excluded middle*. In Harper et al. (Eds.), 87–104.
Starr, W. (2022). Counterfactuals, In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2022 ed.), https://plato.stanford.edu/archives/win2022/entries/counterfactuals/.
Stefánsson, H. O. (2018). Counterfactual skepticism and multidimensional semantics. *Erkenntnis*, *83*, 875–898.
Strevens, M. (2003). *Bigger than chaos: Understanding complexity through probability*. Harvard University Press.
Strevens, M. (2013). *Tychomancy: Inferring probability from causal structure*. Harvard University Press.
Swanson, E. (2012). Conditional excluded middle without the limit assumption. *Philosophy and Phenomenological Research*, *85*(2), 301–321.
Tichý, P. (1976). A counterexample to the Stalnaker-Lewis analysis of counterfactuals. *Philosophical Studies*, *29*(4), 271–273.
Todd, W. (1964). Counterfactual conditionals and the presuppositions of induction. *Philosophy of Science*, *31*, 101–110.
Weatherson, B. (2001). Indicative and subjunctive conditionals. *Philosophical Quarterly*, *51*(203), 200–216.
Wenmackers, S. (2019). Infinitesimal probabilities. In J. Weisberg & R. Pettigrew, (Eds.), *Open handbook of formal epistemology*. PhilPapers Foundation (pp.199–265).
Williams, J. R. G. (2008). Chances, counterfactuals, and similarity. *Philosophy and Phenomenological Research*, *77*(2), 385–420.
Williamson, T. (2020). *Suppose and tell: The semantics and heuristics of conditionals*. Oxford University Press.

**How to cite this article:** Hájek, A. (2025). Similarity accounts of counterfactuals: A reality check. *Philosophy and Phenomenological Research*, *110*, 887–915. https://doi.org/10.1111/phpr.13138

19331592, 2025, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/phpr.13138 by Australian National University, Wiley Online Library on [27/05/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License