Noûs

NOÛS 00:0 (2015) 1–28
doi: 10.1111/nous.12085

Elusive Counterfactuals[*]

KAREN S. LEWIS
Barnard College, Columbia University Department of Philosophy

Abstract

I offer a novel solution to the problem of *counterfactual skepticism*: the worry that all contingent counterfactuals without explicit probabilities in the consequent are false. I argue that a specific kind of contextualist semantics and pragmatics for would- and might-counterfactuals can block both central routes to counterfactual skepticism. One, it can explain the clash between would- and might-counterfactuals as in: (1) If you had dropped that vase, it would have broken. and (2) *If you had dropped that vase, it might have safely quantum tunneled to China*. Two, it can explain why counterfactuals like (1) can be true despite the fact that quantum tunneling worlds are among the most similar worlds. I further argue that this brand of contextualism accounts for the data better than other existing solutions to the problem.

1. Introduction

Suppose a small child picks up an extremely fragile vase from a table in a marble foyer. "Put that down!" Her mother admonishes. After she does, the mother explains:

(1) If you had dropped that vase, it would have broken.

(1) is a *counterfactual conditional*, of the sort we use all the time, and indeed one we take to be true in the described situation. Call the thesis that all such counterfactuals—contingent ones without explicit probabilities in the consequent—are false *counterfactual skepticism*. The threat of counterfactual skepticism comes from many sources. For example, Alan Hájek (ms), David Lewis (1986), and John Hawthorne (2005) worry that extremely low probability events that quantum mechanics tell us are possible make counterfactuals like 2 false. Keith DeRose (1999)

[*] Very special thanks to Alan Hájek for extensive comments and conversations about this paper. Thanks are also due to Chris Barker, Sara Bernstein, Alexis Burgess, John Collins, Janice Dowell, Maya Eddon, Nina Emery, Gabriel Greenberg, Shieva Kleinschmidt, Daniel Nolan, Jason Stanley, Catherine Sutton, and Katja Vogt. Earlier drafts of this paper benefited enormously from the comments and questions of audience members at the 2014 Pacific APA symposium on counterfactuals, the University of Chicago Linguistics and Philosophy workshop, the 2014 Philosophy Mountain Workshop, the Australian National University probability reading group, the 2013 NYSWIP Lectures in Honor of Ruth Barcan Marcus, a 2012 colloqium at California State University-Northridge, the UCLA Language Workshop, and the Coalition of Los Angeles Philosophers. All mistakes are my own.

and Hájek (ms) worry about low probability possibilities that have nothing to do with theories of physics. To these cases, I will add similar ones stemming from even higher probability, more ordinary possibilities.

I'm going to argue that there's good reason to reject counterfactual skepticism and that the right solution is a contextualist semantics and pragmatics for counterfactuals. There are two essential features to the account: first, counterfactuals are not just evaluated relative to the most similar antecedent worlds, but relative also to the worlds *relevant* given the context. Second, might-counterfactuals can change what counts as relevant in a context. These two features are compatible with more than one semantic theory, but I will implement them in the tradition of a Lewis-Stalnaker variably strict conditional semantics.[1]

The next section describes the problem of counterfactual skepticism in more detail. In reading the section, keep in mind that even if you disagree with the claim that *all* contingent counterfactuals without explicit probabilities in the consequent fall prey to the skeptical worries described, a worrisome enough thesis follows from the cases below: at least very many of the counterfactuals we ordinarily take to be true are false. §3 lays out the semantics and pragmatics of both would- and might-counterfactuals, showing how we can reject counterfactual skepticism and account for the puzzle cases that motivate it. §4 defends the contextualist thesis against objections. §5 addresses the question of why we should adopt the contextualist account rather than embrace counterfactual skepticism (and an error theory along with it) or some other purported solution to the problem. Finally, I conclude with a brief consideration of the role of counterfactuals in theoretical contexts, given the contextualist thesis.

2. The Sources of Counterfactual Skepticism

2.1. Quantum Mechanics and Statistical Mechanics

Ordinary speakers in ordinary circumstances take (1) to be true, given the details of the scenario. In fact, if we're sure of any contingently true counterfactual, we're sure of things like (1). The problem is that in an indeterministic universe as described by many interpretations of quantum mechanics, virtually any outcome has *some* chance of happening. And this chance makes sentences like (2) seem true of the same scenario:

(2) If you had dropped that vase, a quantum event might happened in which it flew sideways and landed safely on the couch.

And (2) straightforwardly implies (3):

(3) If you had dropped that vase, it might not have broken.

And (3)—to use DeRose 1999's terminology—*inescapably clashes* with (1), the counterfactual we were so certain was perfectly true. That is, we cannot say (4):

(4) # If you had dropped that vase, it might not have broken, but of course if you had dropped that vase, it would have broken.

But much of our best science indicates that (2) is true. Thus the worry that (1) must be false after all.

The same problem arises even if the causal laws of our universe turn out to not be the indeterministic ones of quantum mechanics, but the regular old deterministic ones of statistical mechanics. The antecedents of counterfactuals like (1) are underdescribed—there are many ways in which the child could have dropped the vase. She could have held it slightly higher or lower, more to the left or the right, with more or less force. The macro-physical description in the antecedent does not specify a micro-physical description of the dropping. And statistical mechanics tells us that while most micro-physical states lead to the unremarkable outcomes we expect—such as the vase dropping—on some micro-physical states, if the atoms line up just so, very weird things happen, such as the vase flying sideways and landing safely on the couch. And so we get exactly the same puzzle we got under indeterminism.

2.2. Underspecification and Low Probability Outcomes

The problems presented above aren't just problems caused by the astronomically unlikely outcomes that quantum mechanics and statistical mechanics tell us are possible. The same puzzle can be generated by looking at ordinary low-probability outcomes. DeRose (1999) raises the puzzle using the following example:

> The score was tied in the bottom of the ninth, I was on third base, and there was only one out when Bubba hit a towering fly ball to deep left-center. Although I'm no speed-demon, the ball was hammered so far that I easily could have scored the winning run if I had tagged up. But I didn't. I got caught up in the excitement and stupidly played it half way, standing between third and home until I saw the center fielder make his spectacular catch, after which I had to return sheepishly to third. The next batter grounded out, and we lost the game in extra innings.
>
> This thought haunts me:
>
> **(5)** If I had tagged up, I would have scored the winning run.
>
> Given the circumstances as I described them above, this is close to being as clear a case of a contingently true counterfactual conditional as one might hope to find. But a skeptic might suggest a little caution here. She puts forward the following conditional, inviting me to agree:
>
> **(6)** If I had tagged up, I might have tripped, fallen, and been thrown out. (DeRose, 1999, p.385, my numbering of examples)

As he points out, (6) incites us to accept (7), which again inescapably clashes with (5):

(7) If I had tagged up, I might not have scored the winning run.

Similar to the statistical mechanics case, this is a problem of underspecification—there are many ways in which the speaker can tag up (his feet, his arms, etc. placed just so or just so), and some of these ways inevitably result in his tripping and falling. These sorts of examples easily multiply—and the outcomes that ultimately undermine the counterfactuals are not always extremely improbable. To borrow a case from Thony Gillies (2007) (which he uses to make a different, though not unrelated, point): suppose Sophie was thinking of going to the parade, but in the end decided not to. We go to the parade, and Sophie's favorite baseball player, Pedro Martinez, performs a dance on a float. Later on, lamenting Sophie's absence, I say to you:

(8) If Sophie had gone to the parade, she would have seen Pedro dance.

This is the sort of thing that native speakers generally accept as true in the situation described. But of course, given all the things we know about Sophie and the parade we were at—there were some tall people standing in the front, Sophie is of normal height, Sophie doesn't tend to wear stilts or push everybody out of her way — you might reply with the following:

(9) If Sophie had gone to the parade, she might have been stuck behind someonetall (and so not seen Pedro dance).

(9) undermines our confidence in the truth of (8) in the same way as the examples in the previous cases. The possibility of Sophie ending up behind someone taller than her is not an astronomically unlikely, esoteric possibility. But we find counterfactuals like (8) accepted all over the place in conversation.

The clash between woulds and mights suggests to many that at least one of the following must be given up:[2]

1. Would counterfactuals like (1), (5), (8) are true.[3]
2. Might counterfactuals like (2), (6), (9) are true.
3. Would and might counterfactuals are duals.

By contrast, on my preferred theory of counterfactuals, we can maintain all three of the above, all the while explaining the clash between woulds and mights. But before I turn to the view, I want to point to another source of counterfactual skepticism, one that is not motivated by the clash between woulds and mights. Any solution to the problem of counterfactual skepticism must address *both* of these routes to the conclusion.

2.3. Not (Just) a Problem of 'Might's

Counterfactual skepticism doesn't just come about because of a clash between mights and woulds; (at least a large portion of) the problem can be motivated without talking about mights at all—it has to do with the similarity ordering crucial to the semantics of counterfactuals.[4] I will argue that on a natural construal of a similarity ordering, the possibilities that underwrite the truth of the pesky might-counterfactuals are among the most similar worlds, and so if the truth of would-counterfactuals is sensitive to all the closest worlds, they are not true in the first place.

Adopting the limit assumption and suppressing the uniqueness assumption, on the Lewis-Stalnaker semantics for counterfactuals a would-counterfactual $P \Box\!\!\rightarrow Q$ is true iff all the closest P-worlds are Q-worlds, where closeness is a function of similarity.[5] A natural construal of the similarity relation, following Lewis, Stalnaker, and Bennett, is as follows. The most similar worlds are those that perfectly match the actual world (α) until some time not long before the antecedent time (following Bennett 2003, call this the time of the fork, T_F), when some small, local violation of the laws of α causes the world to diverge such that the antecedent occurs. The laws of α hold thereafter.[6] By these criteria, the undermining outcomes noted above are all among the closest worlds. Consider worlds that are just like ours until around the time the little girl is holding the vase. Then a small, unoticeable miracle happens—some particle is in a slightly different place, some neuron fires differently—resulting in her dropping the vase. Then the actual laws do what they will given the physical facts in that world. By hypothesis, the actual laws are the laws of quantum mechanics, which predict among such worlds will be ones in which the vase lands safely. Similarly, consider Sophie and the parade. There is more than one small change that will cause Sophie to get to the parade, given that the precise way in which she is at the parade is underspecified by the antecedent. Let the laws play out in the various worlds just like ours except for the small changes made at T_F. These worlds will result in Sophie being in slightly different positions at the parade. There are many such worlds, since there are many minimally different ways in which she could end up at the parade, and given our assumptions about the parade, some of these are going to result in her having a blocked view.

There's one more element to this argument: I haven't mentioned anything about whether match of particular fact after the time of the fork counts for anything. Lewis famously held that"it is of little or no importance to secure approximate similarity of particular fact" (Lewis, 1986, p. 48) and that different cases come out differently—some go wrong if we don't hold steady some particular fact after T_F; others go wrong if we do. For example, if Ann tosses a fair coin and it lands heads, we do not think that the following counterfactual is true: *If Barbara had tossed the coin instead of Ann, it would have landed heads*, but we do think this is true: *If Barbara had bet on heads, she would have won*. Bennett (2003) argues that match of particular fact after the fork counts for closeness when it is part of the same causal chain as in the actual world, but when the antecedent disturbs the causal chain, match of particular fact counts for nothing. This seems to get it right when it comes to our intuitions about particular cases. Since the cases from §2 are all cases in which the antecedent changes the causal chain leading to the consequent, particular fact after the time of the fork should count for nothing.

In any case, match of particular fact after T_F, even if it did count towards closeness, doesn't help us here. In the parade and baseball cases, matching particular fact yields undesirable results: worlds in which the ballplayer still loses the game and ones in which Sophie doesn't see Pedro dance better match α in particular fact than worlds in which he wins and she sees Pedro. And they don't help that much with the physics cases either—dependng on how we count particular facts, it looks like we have a tie. For example, quantum vase worlds mismatch the actual world in having an astronomically improbable quantum event happen, but match insofar that the vase remains unbroken. So match of particular fact doesn't help us here, and is in general a dangerous thing to appeal to (recall Kit Fine 1975's famous Nixon example).

In sum: on a natural interpretation of a Lewis-style similarity ordering, the vase flying sideways and landing safely on the couch, the baseball player tripping and falling, and Sophie getting stuck behind someone tall at the parade count as among the closest worlds. This is not to say that it is impossible do something fancier to the similarity ordering to make these worlds end up farther away—this is the strategy of Lewis (1986) in invoking quasi-miracles and Robert Williams (2008) in invoking typicality. But it is to say that this is another pressure towards counterfactual skepticism. A solution to the problem has to address on one side the clashes between woulds and mights and on the other the fact that the similarity ordering includes ¬Q-worlds among the closest P-worlds.

3. The Semantics and Pragmatics of Counterfactuals

Considering both the matter of the clash and the problem with the ordering, I think this constitutes good evidence that counterfactuals are context-sensitive in a way not generally recognized in the literature.[7] Counterfactuals do not just quantify over the most similar worlds, but the worlds that are *relevant in the context*, or, roughly speaking, the semantic content of a counterfactual is sensitive to the standards of precision in the conversation.

A useful comparison here is the case of absolute gradable adjectives like *flat*. If we take *flat* to mean a complete absence of bumps, divets, and undulations, then basically no objects are flat. We could take this as a reason to think that when someone says *4th Avenue is flat* they are strictly speaking uttering a falsehood (this is what Unger 1975 concludes). But a better explanation—the one I think semanticists and philosophers of language have tended to agree upon—is that the semantic content at a context of a sentence containing *flat* depends on some sort of contextual parameter, such as a comparison class or standard of precision. An utterance of *4th Avenue is flat* might be true in a context in which we want to go for a long, easy bike ride—where small bumps in the road can be legitimately ignored—whereas an utterance of the same sentence may be false in a context in which we're looking for a surface for our very large physics experiment—small bumps cannot be legitimately ignored here. I contend that similarly, when it comes to counterfactuals, in some contexts, we can legitimately ignore possibilities that cannot be legitimately ignored in other contexts.[8]

The central insights of this theory of counterfactuals are twofold. First, counterfactuals quantify over not only the most similar worlds, but over *relevant* worlds, which are not generally co-extensive with the most similar worlds. Second, might-counterfactuals can, through their pragmatic effect, *expand* the domain of relevant possibilities. These idea are compatible with more than one kind of semantics for counterfactuals. I will implement them in a way that stays more or less close to the Lewis-Stalnaker variably strict conditional, because I think this is a nice way to do it. But it is by no means necessary to maintain a Lewis-Stalnaker semantics to solve the problem of counterfactual skepticism. If it turns out that either a Kratzer-style semantics, or a dynamic semantics, or a causal model semantics is correct, these insights can be incorporated into those frameworks. I also want to remain neutral on what sort of context-sensitivity this is, for example, I am not making any claims on whether the context-sensitivity is realized at the level of syntax. The only claim I am endorsing is that it is a *semantic* context-sensitivity, in the sense that the same words can be used in one context to express a true proposition and in others to express a false one (as opposed to some sort of pragmatic context-sensitivity, in which the same proposition is always strictly speaking semantically expressed, but can pragmatically convey different propositions in different contexts).

3.1. Would-Counterfactuals

There are (at least) two ways in which relevance can be incorporated in the variably strict semantics, and there are benefits and downsides to each. Since both yield the right results for present purposes, I will leave deciding between them for future work. The first way is to maintain the classic Lewis-Stalnaker semantics, but add a parameter for relevance to the ordering source:

> Option 1: For all contexts c, $P \Box\!\!\rightarrow Q$ is true in c iff all the closest P-worlds are Q-worlds, where closeness is a function of both similarity and relevance.

The second option keeps a single ordering source, based on similarity, but changes the basic Lewis-Stalnaker semantics:

> Option 2: For all contexts c, where d is the relevant subset of of the most similar P-worlds, $P \Box\!\!\rightarrow Q$ is true in c iff all the P-worlds in d are Q-worlds.

As I said above, as far as I can tell, the two options yield the same results when it comes to the sort of cases I'm concerned with in this paper. But there are other differences that may ultimately decide between the two. Option 2 is only successful insofar as the largest domain that counterfactuals (both woulds and mights) are ultimately sensitive to is the set of most similar worlds. If there is reason to think that counterfactuals can be sensitive to farther away worlds (either because of evidence from certain might-counterfactuals, or reverse Sobel sequences), option 2 fails. Option 1 is nice in that it maintains the classic semantics, but has the downside of requiring 2 ordering sources to work together. There is good reason to think this is possible though. First, some work has been done on merging two ordering sources in the domain of deontic modals.[9] Second, the idea that relevance and similarity could work together in inducing an ordering has some precedence in the idea that sometimes certain kinds of differences count as negligible for the purposes of the conversation; for example, if we're considering worlds in which this room is smaller, worlds in which this room is anywhere from one square inch to one square foot smaller might count as equally close (see §4 below for further discussion of this point). Picturesquely speaking, relevance takes the similarity ordering, and squishes some worlds that were in farther domains into the closest domain, and pushes others that were closer farther out. On the other hand, another downside to option 1 is that changes in relevance induce a *reordering* on worlds, while changes in relevance on option 2 merely induce a domain expansion. This makes option 1 less elegant, but not by any means unworkable.

What does it means for a possibility to be relevant? Or, to ask the same question in the inverse way, what does it mean to legitimately ignore a possibility? The main idea is that the relevant possibilities are the ones that fulfill the current conversational purpose, for example, warning a child, expressing regret, lamenting a friend's absence at a parade. While speakers' intentions play a role insofar as they can partially determine the conversational purpose, I do not think speakers have the power to determine the domain of worlds over which counterfactuals quantify. I'm thinking of contexts and contextual parameters, in this sense, as objective features of the conversation. Counterfactuals are generally used to make predictions (and sometimes dispositional claims). What falls out of this observation? If counterfactuals are used to make predictions (of various standards of precision), then the actual world matters. Similarity to the actual world matters because predictions about far away worlds aren't important to us. And if the actual world is among the closest antecedent worlds, then it is always relevant; if one is trying to predict what would have happened given P, then what did happen given P is relevant. High probability possibilities are also always relevant (though how high is high is vague), while low probability events can sometimes be ignored, depending on the specifics of the context. Here "high" and "low" apply to macro-physical descriptions of events and not micro-physical descriptions, since low probability things (at the micro-physical level) happen all the time (e.g. an example of a macro-physical description is "the vase breaks"; an example (of a schema) of a micro-phsyical description is "the vase breaks in precisely such and such a way, with the pieces falling precisely here and there"). Note that unlike Lewis (1986)'s quasi-miracle, which is both a low probability *and* a remarkable event, my notion of relevance has nothing to do with remarkableness in general (though of course, there may be particular contexts in which remarkableness plays a role).[10]

For example, in the vase case in §2.1, the mother's conversational purpose is clearly to admonish her daughter and teach her a lesson about how to treat fragile objects in the future. 2 expresses a dispositional property of the vase the daughter was holding, and indeed provides a useful piece of information about it. In these sorts of everyday endeavors, astronomically low probability events just aren't relevant. It is enough to appeal to worlds in which things go how our world generally goes—even if the laws of our world allow for exceptions to these generalizations.

To take another example, consider Sophie and the parade again. Suppose we're talking about how awesome it was to see Pedro dance at the parade. We lament that our friend Sophie, who wasn't able to come to the parade, missed Pedro's dance. In this context, an assertion of (10) is true:

(10) If Sophie had come to the parade, she would have seen Pedro dance.

This is true even though we both know parades in general have, and that this parade in particular had, a mass of disorganized people of various heights and that Sophie is of average height.[11]

Suppose, by way of comparison, that we are talking about how parade watchers lack etiquette and we wish there was some system to ensure that everybody who comes to the parade gets to see the parade. In this context, an assertion of (10) is false *even thoguh all the facts about the parade are the same as the other context*.

Note that on this view, the intersubstitutivity of logical equivalents does not hold (for the antecedent or consequent), since more precise formulations of an equivalent proposition—e.g. the use of a disjunction that employs micro-physical descriptions to express all the different ways in which a vase could be dropped—will land us in a highly unordinary, precise context, while the logically equivalent *you dropped the vase* will not.

3.2. Might-Counterfactuals

I want to maintain that might-counterfactuals are, semantically, the duals of woulds, that is:

> A might-counterfactual $P \Diamond\!\!\rightarrow R =_{def} \neg(P \Box\!\!\rightarrow \neg R)$, i.e., For all contexts c, a might-counterfactual $P \Diamond\!\!\rightarrow R$ is true in c iff some closest P-world is an R-world, where closeness is a function of similarity and relevance (Option 1)/some P-world in d is an R-world, where d is the subdomain of relevant P-worlds (Option 2).

I have been arguing that in the vase, baseball, and parade cases from §2, worlds in which quantum events happen, DeRose trips, and Sophie is stuck behind someone tall aren't relevant in the contexts in which the would-counterfactuals are uttered. It follows that might-counterfactuals like (2), (6), and (9) are *false* in those same contexts. So how to explain their apparent truth, and the ensuing clash with the woulds? I want to adopt an insight from Lewis (1979) and Gillies (2007) that mights can induce a context *shift*.[12] Taking the vase case for example, (2) shifts the context to include worlds in which quantum events occur that were previously legitimately ignored. (2) is true in the updated context. Though (1) was true in the context in which it was originally uttered, it is false in the new, updated context. The idea is that when a possibility is introduced that appears to undermine the prediction being made with a counterfactual—as in the case of (1) and (2)—and it cannot be ruled out, either as not a possibility or as irrelevant, it *becomes* a relevant consideration. That mights can introduce new possibilities is an observation that traces back to Lewis (1979), where he describes the following case:

> Suppose I am talking with some elected official about the ways he might deal with an embarassment. So far, we have been ignoring those possibilities that would be political suicide for him. He says: "You see, I must either destroy the evidence or else claim that I did it to stop Communism. What else can I do?" I rudely reply: "There is one other possibility—you can put the public interest first for once!" That would be false if the boundary between relevant and ignored possibilities remained stationary. But it is not false in its context, for hitherto ignored possibilities come into consideration and make it true. (p. 354-5)

Consider another example. Suppose a group of teenagers, who were all grounded, are discussing how their evening would have gone if they had not been grounded.

Since they all share a love of movies, so far the conversation has centered on which movie they would have seen. They might even come to agree on something like the following:

(11) We would have gone to see *American Reunion*.

That is, they might all agree on (11) until someone says:

(12) We might have gone to Jane's party.

Until the point in the conversation in which (12) is uttered, no party-going worlds are among the possibilities they were reasonably entertaining. That is, there were no party-going worlds relevant to their modal talk. But assuming the teenagers don't find a reason to reject the utterance, now that (12) has been asserted, the party-going possibility must be taken into consideration.

There are at least two reasons for thinking this role unembedded *might* plays extends to might-counterfactuals. First, there's good reason to think the *might* outside of a might-counterfactual and the *might* within one are related, in that they have the same meaning and can play the same roles. This is demonstrated by the above example involving the teenagers: there's no reason to think that *might* can play the role it does in (12) but not if they had made the counterfactual reasoning explicit and said *If we had not been grounded, we might have gone to Jane's party*. Second, it does justice to the data. I've been arguing that normal use of counterfactuals reveals that would-counterfactuals are sensitive to only a relevant subset of the most similar worlds. If we also want to maintain that our intuitions are correct about when might-counterfactuals are true, then we have to accept that might-counterfactuals can be sensitive to a different (larger) domain than would-counterfactuals.

As I mentioned above, the change in context induced by the might-counterfactual is a pragmatic one. In the context as it stands after the utterance of the initial would-counterfactuals in the vase, baseball, and parade cases, a speaker of one of the undermining might-counterfactuals has said something false. To account for the fact that speakers don't generally assert obvious falsehoods, interlocutors have to accommodate the fact that the speaker is taking heretofore (legitimately) ignored possibilties to be relevant. In other terms, in uttering a might-counterfactual (that involves previously ignored possibilities), the speaker raises the conversational stakes—she takes lower probability events to be relevant. The conversational participants accommodate by adding the appropriate worlds to the domain (either by reordering worlds or by expanding the domain of relevant worlds depending on whether we adopt option 1 or 2). It is relative to this updated context that the might-counterfactual is true (this is no different from other typical cases of accommodation).

This is not to say that the conversational participants *must* accommodate and change the context. They can challenge the truth or the relevance of the might-counterfactual; this is certainly a likely response in the quantum cases. I also want to note that I am not arguing that all might-counterfactuals (or even all might-counterfactuals relative to a particular domain e.g. those that don't violate certain laws) are made true by uttering them.[13] Some are just false and should not be accommodated, because the possibility is too far away (and thus also determinately irrelevant). I'll say a bit more about mights and accommodation in the next section.

3.3. Putting the Pieces Together

We are now in a position to see how the pieces fit together to solve the puzzles from §2.

CASE 1: VASES AND PHYSICS

Recall the examples from the case:

A mother says to her daughter, who had picked up a fragile vase,

(13) If you had dropped that vase, it would have broken.

But our scientifically-minded, skeptical friend comes along and says:

(14) If you had dropped that vase, it might have flown sideways and landed safely on the couch.

This immediately implies:

(15) If you had dropped that vase, it might not have broken.

And the latter inescapably clashes with (13):

(16) # If you had dropped that vase, it might not have broken; but if you had dropped that vase, it would have broken.

According to the proposed analysis, (13) is evaluated relative to the relevant similar worlds—worlds that are like the actual world in the ways described by a classical similarity-relation, but only those in which things go as they typically go. Quantum events aren't relevant to the conversation. (14) clearly takes quantum events to be relevant. These quantum possibilities are accommodated, making (14) true in the updated context, since among the relevant similar worlds are now worlds in which quantum events occur. And this explains the clash: once we've taken the might-counterfactual and the resulting expanded domain on board, the would-counterfactual is no longer true in the updated context, since the relevant similar worlds include some vase-and- ¬break-worlds.

CASE 2: BASEBALL AND REGRET

Recall the key examples from the DeRose case:

It's the bottom of the ninth with a tied score and the ball is hit deep into left field. DeRose is on third base and has ample time to tag up and score the winning run, even if he runs at a mediocre pace. But he gets caught up in the excitement and doesn't tag up, and his team loses in overtime. He immediately regrets his inaction, thinking:

(17) If I had tagged up, I would have scored the winning run.

But the skeptic cautions him by pointing out that he might have tripped and fallen, and so he accepts:

(18) If I had tagged up, I might have tripped, fallen, and been thrown out.

Again, this implies:

(19) If I had tagged up, I might not have scored the winning run.

With (17) the speaker expresses regret—expressing regret is one common purpose of counterfactuals. In expressing his regret (given the facts about the situation), low probability events in which he fails anyway just aren't relevant. DeRose's skeptic in uttering (18) clearly takes even low probability events in which the speaker fails no matter what he does to be relevant; in the expanded domain (18) and (19) are true and (17) is false.

The story for the parade case goes in much the same way. Recognizing that the semantics of counterfactuals are sensitive to relevant possibilities and that might-counterfactuals are context-shifters allows us to have our cake and eat it too: in everyday contexts, ordinary utterances of counterfactuals are true. Thus counterfactual skepticism is avoided. Might-counterfactuals and would-counterfactuals are still duals, but since might-counterfactuals are context-shifters, the clash is explained—the mights and woulds really do contradict each other—without falling into counterfactual skepticism.

4. A Defense of Contextualism for Counterfactuals

It is not a new claim that counterfactuals are context-sensitive. The classic Lewis-Stalnaker semantics incorporates context-sensitivity in at least two ways. First, there are cases like the famous Caesar in Korea example, where depending upon whether we are talking about an ancient Caesar being transported to modern times without knowledge of modern weaponry, or whether we're talking about Caesar being born or at least educated in modern times, one of the following will be true and the other false (but they are not both true in the same context):

(20) If Caesar had been in command in Korea, he would have used the atom bomb.
(21) If Caesar had been in command in Korea, he would have used catapults.

The kind of context-sensitivity at hand is different, because it is context-sensitivity *after all the facts are fixed.* Fix whether we're talking about ancient or modern Caesar. Counterfactual skepticism can still raise its ugly head in the form of an undermining might-counterfactual.

The second kind of context-sensitivity, mentioned briefly in §3.1 above, is discussed in Stalnaker (1968, 1981) and Bennett (2003). It regards what counts as a negligible difference when it comes to minimal or maximal units of measurement.

For example, supposed there are 20 people packed tightly into a small room at a party and someone says:

(22) If this room were any smaller, we wouldn't all fit in here.

Clearly the speaker is not considering cases in which the room is 10^{-10} cm smaller. Perhaps the relevant minimal unit here is one square inch. But we don't think it's true that if the room were any smaller, it would be exactly one square inch smaller. Maybe the relevant interval is anywhere from a square inch to a square foot. All these worlds come out equally close; the difference in size doesn't matter for closeness. This is getting closer to the sort of context-sensitivity I have been arguing for, since it is normally thought that relevance determines the range of negligible differences. But the sort of examples I've been concerned with don't have to do with negligible differences. Relevance in these examples plays a different role (though it is possible that my view subsumes this other role that relevance plays). Furthermore, the negligible size differences may also have to do with the context-sensitivity of the specific measurement terms involved (rather than the counterfactual construction itself)—if we are at a publishing company and I say "make the margins bigger" and you make them 10^{-10}cm bigger, there is a clear sense in which you have not followed my instructions, or at least brazingly flouted my intended meaning.

The brand of context-sensitivity I am arguing for is more extreme; it is a brand under which the truth of a counterfactual is fragile, highly unstable across shifts in conversational context. And while (arguably) the sort of existing context-sensitivity posited for counterfactuals is not susceptible to traditional arguments against contextualism, the sort I have been arguing for is (particularly those trotted out against contextualism for knowledge claims). So it is a defense of my contextualism for counterfactuals against these sorts of objection that I now turn to.

Of course, some people (semantic minimalists) think there is very little context-sensitivity in natural language—perhaps only the automatic indexicals are instances of semantic context-sensitivity. I will have little to say to this sort of objection, simply because there is no space to give a complete defense of semantic context-sensitivity here. Rather, my strategy (following Stanley 2005's argument against, and DeRose 2009's argument for, contextualism for *know*) will be to give an inductive argument that counterfactuals share important features with expressions that are better established as context-sensitive, and do not face any problems that at least some of these other expressions don't. The aim here is to show that positing contextualism for counterfactuals is not unwarranted by the linguistic data, nor does it create any *new* problems.

There is no single feature (aside from context-sensitivity) that all context-sensitive expressions share and context-invariant expressions lack. Furthermore, I contend that any two categories of context-sensitive expressions differ in some important way. Thus, there is no *one thing* that is context-sensitivity. Context-sensitivity is more like family resemblance than a single semantic property. And it may be that different cases should be treated quite differently at the syntactic and semantic level. All I want to show is that counterfactuals belong in this family.

Here are some categories of context-sensitive expressions:

1. Automatic indexicals: *I*, *you*, *today*, etc.
2. Gradable adjectives: *tall*, *rich*, *boring*
3. Absolute gradable adjectives: *flat*, *empty*, *pure*
4. Quantifiers: *All the beer*, *every bottle*, *many students*
5. Meteorological predicates: *raining*, *snowing*, *humid*
6. Determiners: *many*, *most*
7. Relational expressions: *enemy*, *local*, *coming* and *going*
8. More controversial cases:
   (a) Knowledge: *Bob knows that he has hands*
   (b) Predicates of taste: *good*, *tasty*, *pretty*
   (c) Epistemic modals: *That might be Bob approaching*

I will not rest my defense of contextualism for counterfactuals on any of the controversial cases, only mentioning them when it is directly relevant.

The objection I have received most often, a version of which Stanley 2005 levels against contextualism for *know*, is an objection based on the impossibility of various sorts of propositional anaphora across contexts. A contextualist theory, including the one I have been arguing for, predicts that speakers or other conversational participants should be able to judge as true what was said in an earlier context, even though the same words don't express a truth in the current context. Furthermore, they should be able to use propositional anaphora to *call* true the sentence that is predicted to be true by the theory in its context, even though the current context (the one in which the propositional anaphor is being used) is not one in which the same words express a truth. Consider the following examples. Objectors claim that the first three discourses involving context-sensitive expressions are felicitous. But the corresponding discourses involving counterfactuals, though each sentence is predicted to be true by my theory, are infelicitous. (The first two examples are from Stanley 2005.)

(23) A: It's possible to fly from London to New York City in 30 minutes.
B: That's absurd! No flights available to the public today would allow you to do that. It's not possible to fly from London to New York City in 30 minutes.
A: I didn't say it was. I wasn't talking about what's possible given what is available to the public, but rather what is possible given all existing technology.

(24) It's raining here. Had I been inside, what I said still would have been true. But now that I am in fact inside, it is not raining here.

(25) A: (*In New York, on the phone with B*) It's raining here.
B: (*In Paris*) That's true, but I'm in Paris, so it is not true that it is raining here.

(26) A: If you had dropped that vase, it would have broken.
B: But our best physics says that there is a chance that anything could happen, for example, the vase could quantum tunnel to China and land safely. So it's not true that if you had dropped that vase, it would have broken. It might have landed safely.
A: True.

B: So you admit that you were wrong when you said that if I had dropped the vase, it would have broken.
A: ?? I didn't say that it would. I was only making a claim about what would have happened had things gone *normally*.

(27) If I had dropped that vase, it would have broken. But come to think of it, there's a chance that the vase might have quantum tunneled to China and landed safely, in which case it wouldn't have broken. ?? But what I said earlier is still true.

(28) A: If you had dropped that vase, it would have broken.
B: ?? That's true, but there's a chance of it quantum tunneling to China and landing safely, in which case if you had dropped the vase, it wouldn't have broken.

Note that the objection is not that there are no felicitous discourses in the vicinity of (26), (27), and (28). There are small changes we can make to the discourses to make them felicitous, but they mostly involve removing the propositional anaphora. Rather, the claim is that this is evidence against a semantic treatment of context-sensitivity, as on the latter view, we *should* be able to use propositional anaphora to pick out true propositions expressed in a different context.

My reply to the objection takes two parts. First, I will argue that analogous discourses involving other context-sensitive expressions are also infelicitous.[14] Second, I will offer a (partial) diagnosis of the infelicity. My first claim is that if the above is good evidence against contextualism for counterfactuals, it is equally good evidence against contextualism for absolute gradable adjectives and quantifiers (as well as epistemic modals and predicates of taste, but as I mentioned earlier, I do not want to rest my case on these examples):

(29) A: That field is flat.
B: But the field has many small divets and bumps. So it's not flat.
A: ?? I didn't say it was. All I was saying is that it was flat for a football field.

(30) That field is flat. But come to think of it, there are of course many small divets and bumps throughout the field, so it's not flat. ?? But what I said earlier is still true.

(31) A: That field is flat.
B: ?? That's true, but there are many small divets and bumps throughout the field, so it's not flat.

(32) A: Everything in the fridge is edible.
B: But the shelves aren't edible. So it's not true that everything in the fridge is edible.
A: ?? I didn't say it was. I was only considering the food items.

(33) Everything in the fridge is edible. Of course, come to think of it, the shelves are not edible. So it's not the case that everything in the fridge is edible. ?? But what I said earlier is still true.

(34) A: Everything in the fridge is edible.
B: ?? That's true, but the shelves are not edible, so it's not the case that everything in the fridge is edible.

Many people want to treat quantifier domain restriction and absolute gradable adjectives as species of semantic context-sensitivity. So the problem is not one that counterfactuals alone face among purported context-sensitive expressions. But the weirdness calls out for an explanation. I think different examples warrant different explanations. Examples like (27), (28), (30), (31), (33), and (34) involve what looks like *genuine disagreement*, either between speakers or between one's present and earlier self. And one is not going to say "that's true" in a case of disagreement—as a matter of fact, the speaker seems to think that what she said or what the other person said originally is *false*. Now this pushes the question back rather than solving it, because why should we expect disagreement given that the theory says both sentences are true? I'll address this shortly, in responding to the next objection. The point here is that discourses like these become a matter of explaining disagreement and retraction phenomena.

The really puzzling cases are the cases like (26), (29), and (32), in which the speaker recognizes that her interlocutor's disagreement is misplaced and that they can both be right, since they are considering different contexts. It's interesting to note that while these discourses are bad, replacing the "I didn't say it was" locution—which picks up on the interlocutor's proposition—with a "what I said was true because" locution—which picks up on the speaker's proposition—much improves the discourses. I admit that this is somewhat mysterious. I suspect it has something to do with the inexactness of using propositional anaphora as locutions like "I didn't say it was" and even ones like "I didn't assert that" can pick up on the words used or the proposition expressed, and so when the same words are used to express different propositions in close successsion, clarifying the context explicitly is preferred. Even Stanley's example is not great—many have reported to me that they do not find it felicitous.[15] This is good evidence that it has something to do with aspects of the discourse that provide little evidence as to whether something is context-sensitive or not.

I just argued that what is weird about many of the cases above is that they appear to be cases of genuine disagreement. In fact, the cases from §2 that motivated this paper seem like cases of genuine disagreement. The skeptic who brings up the undermining might-counterfactual doesn't appear to be merely changing the context, but genuinely disagreeing with the asserted would-counterfactual. This alone is a worry, since given what I've said, both conversational participants have said something true. A related phenomena, to my mind, is that of *retraction*: speakers are tempted to concede to the skeptic and retract their initial statement despite the fact that for all I have said, they can accept the skeptic's challenge and maintain that what they said in the original context was true in that context. (It should be noted that the speaker doesn't *have* to retract what she said. She can refuse to accommodate the skeptic's context shift and insist that what she said was and is true, or she can clarify what she meant by making the context explicit.)

What I mean by (apparent) genuine disagreement is that the disagreement is not easily dissolved by clarifying the context. For example, suppose A and B are having an argument like the following:

(35) A: It is raining!

B: No, it is not raining!

The argument will quickly come to an end if we figure out that though A and B thought they were in the same place, A is actually talking about New York while B is talking about Paris. On the other hand, it doesn't seem like the following argument is assured of quickly coming to an end when A explains that she is only considering the normal course of things. Again, that is not to say that it *can't or won't* come to an end, but that reasonable people can continue arguing after clarification is made. This is quite different from the rain case.

(36) A: If you had dropped that vase, it would have broken.
B: That's not true! A quantum event might have occurred in which it landed safely.

Again, counterfactuals are not alone in this category, so it is not an objection to the present theory alone:

(37) A: This field is flat.
B: That's not true! It has bumps all over the place.

What counts as genuine disagreement doesn't sort into a sharp dichotomy, but rather forms a spectrum. For example, the following discourse is not as clearly resolvable by clarification as the rain case, though perhaps it is more so resolvable than cases involving absolute gradable adjectives or counterfactuals. Consider a case in which A and B are preparing for a party they are throwing:

(38) A: Every bottle of beer is in the cooler.
B: That's not true! Our secret stash of craft brew is still in the basement.

Suppose A clarifies that she was only considering the bottles they planned to serve to guests. This may clear up the argument, or B may continue to insist that what A said was false.

What I think is going on in these cases is not disagreement in the sense of disagreeing about the truth of a specific proposition, but rather disagreement about what the relevant context is or should (or can) be. In (35), it simply makes no sense to disagree about the appropriate context (unless the speaker has done something truly inappropriate, like intend to talk about raining in Paris when she is neither in Paris, nor is Paris being depicted on the TV, in a salient newspaper, etc.). But so long as Paris and New York are conversationally available locations in the conversation, one cannot really dispute the truth of what the speaker is saying after she has clarified what she meant—that it is raining in Paris—on the grounds that she should have somehow taken into consideration New York's weather. On the other hand, it is more reasonable to question whether the speaker should have considered all the bottles of beer in the house when she claimed that every bottle was in the cooler. And even more so in the case of counterfactuals, speakers can reasonably and genuinely disagree on whether, say, quantum tunneling worlds are relevant to counterfactual claims about dropped vases. In sum, while the disagreement may superficially appear to be a disagreement about the truth of a proposition, it is actually one over which worlds are relevant for the evaluation of the counterfactual (i.e. which worlds should be included in the domain).

Given this hypothesis about disagreement, we would expect variation even within certain categories (e.g. not all gradable adjectives will act the same), depending on whether there are well-established conventional contexts. I think this is exactly what we find. For example, it seems unreasonable to disagree with someone who claims that her 5-year-old son is tall (based on the fact that he is in the 98th percentile for height for his age) on the basis that he is shorter than the majority of males. This is because gender and age is a well-established comparison class for tallness. On the other hand, it is much more reasonable to disagree with someone who claims that her neighbor is rich (based on the fact that he is in the 98th percentile for income in their middle class neighborhood) on the basis that he has much less money than Wall Street bankers, CEOs, etc. That is, it is reasonable in many contexts to negotiate about the right comparison class for *rich*, much less so for *tall* (though not impossible—this is a spectrum, again, not a sharp distinction). This idea fits into the bigger picture of what I think is going on in many conversations with strings of counterfactuals: they are *negotiations* about which worlds are relevant.

As for retraction induced by a context-shift, this is also a feature not possessed by counterfactuals alone. In fact, Hawthorne (2004) cites it as one of the features of context-sensitivity more generally. Retraction is not, in fact, a feature of all context-sensitive expressions. For example, the automatic indexicals and meteorological predicates both do not induce retraction—I'm not at all tempted to take back my claim that *I am hungry* or *It's raining* just because the context shifts to another speaker and place when you start talking. But it is a feature of many context-sensitive expressions, like quantifiers, some gradable adjectives, and absolute gradable adjectives. In the retraction cases, what I think is going on is that speakers confuse the fact that P is false in the new context with P being false in the original context. And this is just what we'd expect if what counts as an appropriate context is unclear.

The final objection I'd like to consider is that context-sensitive expressions can normally take on different values within a single sentence or a short discourse (perhaps not always, but at least in some contexts). For example, we can say things like:

(39) That rock is flat and that field is flat.

(40) Every sailor waved to every sailor.

As Stanley (2005) points out with these examples, we can say (39) even though what counts as flat for a rock is different from what counts as flat for a field, and (40) can express something like *Every sailor on the boat waved to every sailor on the shore*. Similarly, context-sensitive expressions can "downshift" over the course of a short discourse. In the following example, also from Stanley, the second occurrence of "every van Gogh" has a smaller domain than the one in the previous sentence.

(41) A: Every van Gogh painting is in the Dutch National Museum.

B: That's a change. When I visited last year, I saw every van Gogh painting, and some were definitely missing.

The question is: can counterfactuals act in these ways as well?

Note that examples like (39) work best when the contextual parameters for each occurrence of *flat* can all be described with a single (non-trivial) description, e.g. "things that are flat for the sort of objects they are" (even if the actual comparison classes are different for each occurrence).[16] It is much harder to hear them as good when the contextual parameters cannot be described in this way. For example, suppose we are looking at our new ping-pong table and I comment that it's flat, unlike our old, warped table. However, I go on to point out, in my physics lab it wouldn't count as flat, since it does have some bumps. It sounds weird to go on and say, even with the context set up so clearly:

(42) ?? At home, this table is flat, but in the physics lab it is not flat.

It is really something like the latter context shift that we'd need for counterfactuals, so it's natural to expect that it's hard, if not impossible, to find counterfactuals with mid-sentence context-shifts. One case in which we could find mid-sentence context-shifts is embedded counterfactuals. In most cases, conversational purposes are such that the context is not going to shift in the course of the utterance of a single embedded counterfactual. But I contend that sometimes the context does just that. Consider the following embedded counterfactual (a possible response to the skeptic):

(43) Trust me—even if some quantum events were to occur in the history of our universe, if you had dropped that vase, it would have broken.

If we go about evaluating (43) without context-shift, then the antecedent tells us to go to the closest (i.e. most relevant similar) worlds in which a quantum event occurs at some time in the history of the world. When we move to evaluate the embedded antecedent, we have to go to the closest worlds to those in which the hearer drops the vase. Since quantum possibilities have been accepted into the context, unless a context-shift takes place, among the vase-dropping worlds will be ¬breaking-worlds in which a quantum event takes place. But (43) is perfectly fine—in fact, it seems like something many of us would say when confronted by the skeptic. Thus I contend that this is a case of mid-sentence context shift.

I think counterfactuals can also down-shift over the course of a short discourse. To see this, we need to distinguish between forced and natural contexts shifts. In forced shifts, those induced by the skeptic, if the skeptical possibility is accommodated, it is difficult to downshift to a less strict context. This is true for the case of counterfactuals, as well as that for absolute gradable adjectives and quantifier domain restriction. But in natural shifts—such as the one in (41)—downshifting is no big deal. I contend we can have similar natural shifts in discourses involving counterfactuals. For example, suppose some physicists are chatting in the lab. Physicist B is holding physicist A's favorite cup and A says:

(44) Be careful with that cup! If you drop it, it would most probably break. It might even quantum tunnel to China! That reminds me—I almost dropped the box of new lab equipment when I was bringing it into my office last night. If I had dropped it, I would have been in a lot of trouble!

In this case, A is taking quantum possibilities seriously in asserting the first counterfactual, about the cup. But she naturally ignores quantum possibilities in which the lab equipment, though dropped, lands safely, when she asserts the second. And this seems perfectly fine, analogous to the case of quantifier domain restriction.

Finally, contextualism for counterfactuals can explain why some people do not find that examples like those in §2 inescapably clash—a minority of informants report that they sound just fine. Just like one can keep two contexts in mind at the same time for other context-sensitive expressions—there is a sense in which I can understand how this table is both flat and not flat, how Timmy is tall and not tall—it should be possible to keep in mind two contexts for the counterfactual discourses in §2, and hear both the would-counterfactual and the undermining might-counterfactual as true at the same time.

5. Comparison with (Some) Other Theories

Contextualism is clearly not the only possible response to the problem of counterfactual skepticism (nor is it by any means the only actual response). I think contextualism is the best response, in no small part because it vindicates speakers' intuitions and explains a lot of the data. I do not have time to address rival views in all the detail they deserve. Overall, my general complaint about the other views is this: aside from error theories, which do not aim to solve the problem, all the other views solve one or the other of the problems that motivate skepticism, but not both. That is to say, some explain the clash between mights and woulds. Others explain why would-counterfactuals can be true in the first place. But none of them explain both, while contextualism does.

5.1. Quasi-Miracles and Typicality

David Lewis's own response to this sort of problem was to invoke the notion of a *quasi-miracle*, which is a remarkable, low-probability event. On this version of Lewis's similarity ordering, the presence of quasi-miracles, like miracles, locate a world farther away than one with no or fewer quasi-miracles. Since things like dropped vases flying sideways and landing safely on the couch or quantum tunneling to China are both remarkable and low-probability, they are quasi-miraculous, and thus not among the closest worlds. If this is right, counterfactual skepticism is saved, at least in the face of quantum mechanics (this is the sort of skepticism Lewis was worried about). The relevant would-counterfactuals come out true, and the mights false. Lewis also had a story about why the might-counterfactuals seem true. He thought they were indeed true, on a different reading of the might-counterfactual, one that is not the dual of would, and one that is compatible with would: the "would-be possible" reading: If $P \Box\!\!\rightarrow \Diamond Q$.

Aside from the (already worrisome) fact that being quasi-miraculous involves the pyschological property of remarkableness, quasi-miracles have been taken to task by Hawthorne (2005), whose arguments were largely upheld under scrutiny by Williams (2008). I will not rehearse their arguments against quasi-miracles here, and refer interested readers to their work. I will just briefly add that quasi-miracles do not offer a general solution to the problem, since they do nothing to help in DeRose cases or parade cases, since in the cases described, there is nothing remarkable about tripping while playing baseball or being stuck behind someone tall at a parade.

Williams argues that instead of quasi-miracles, what a Lewisian should invoke is *typicality*. Typicality is a global, holistic feature of an outcome, rather than a feature of the specific micro-physical description of that outcome. It looks "not at the probability of a *particular* outcome arising, but at the probabilities of a suitable set of *properties* which that outcome instantiates" (Williams 2008, 409). So in the case of the dropped vase, falling to the ground and breaking are high-probability properties (even though there are ever so many low probability ways in which this property can be instantiated) while quantum tunneling to China is a low probability property. Thus, if the similarity relation is sensitive to typicality as Williams suggests, quantum tunneling worlds are atypical, and thus farther away than worlds in which the vase falls to the floor. Thus we get the same result as when we invoke quasi-miracles, without invoking remarkableness or incurring any of the other difficulties that quasi-miraculousness raises.

I do not have the room for a complete evaluation of typicality here, and I have no *prima facie* problem with its invocation in the case of some counterfactuals—perhaps it is a nice way to formally cash out what makes quantum outcomes irrelevant in ordinary circumstances. However, I do not think that adding typicality alone to the Lewis semantics—or quasi-miracles either, for that matter—is enough to give a complete account of counterfactuals. First of all, it is not clear that typicality fairs any better than quasi-miracles when it comes to the DeRose cases or parade cases. Typicality, if it is not to bear all the same problems as the quasi-miracle, has to be cashed out in some way that doesn't invoke remarkableness. Following Gaifman & Snir (1982) and Elga (2004), Williams take a typical outcome to be an *objectively random* one. An outcome is random if it has all of the appropriately simple high-probability properties (where what counts as appropriately simple is not exactly clear). In the case of worlds infinite in extent, a random world is one that satisfies all the simple probability-1 properties. Williams does not offer an extension of this to the case of finite worlds, though is optimistic that there is one. In any case, we need not worry about the finite case here, because infinite worlds present enough of a problem for present purposes. This is because while this may work well for sequences of coin flips or quantum events, it does not work so well for more ordinary outcomes, like tripping at a baseball game or getting stuck behind someone tall in the parade. Some (infinite) worlds that contain such events will be among those that satisfy all simple probability-1 properties. And so they will not count as atypical worlds (or local atypicalities, therefore), and are thus still among the closest worlds (given all the other facts already established in the DeRose and parade cases).

Furthermore, even if a version of typicality or quasi-miraculousness could work to explain why the skepticism-inducing outcomes are actually among farther out worlds, such a theory still incurs the following costs vis-a-vis the data. First, it is forced to give up the duality of woulds and mights (at least in these cases) to explain the truth of the might-counterfactuals. Two, it needs an additional story to explain why the would and might counterfactuals seem to inescapably clash (presumably this will be some story about confusing the *would-be-possible* reading of the might-counterfactual with the *not-would-not* reading). Finally, it does not explain the retraction data, or why when atypical outcomes are salient, the relevant would-counterfactuals are judged to be false by competent speakers of the language.

5.2. Probability Theories

Another anti-skeptical strategy is to treat the semantics of would-counterfactuals as invoking something weaker than a necessity modal, specifically:

> $P \Box\!\!\rightarrow Q$ is true iff the vast majority of closest P-worlds are Q-worlds (where the vast majority is some, possibly vague, cut-off point).

That is, $P \Box\!\!\rightarrow Q$ is true iff it is *highly probable* that Q is the case, given P (where what counts as highly probable is some, possibly vague, high probability). This strategy is suggested with some sympathy by Bennett (2003), and discussed and rejected by Hawthorne (2005). The approach faces at least two problems: first, it can't both maintain that woulds and mights are duals and explain the inescapable clashes. Second, as others have pointed out, it has to give up agglomeration (defined below), which is an extremely intuitive principle. This is a high cost.

While the probability theory offers an explanation as to why many ordinary would-counterfactuals are true, it faces a choice-point when it comes to how to treat might-counterfactuals. Neither choice explains all the relevant data. If the probability theorist wants to maintain that woulds and mights are duals, then it becomes more difficult for might-counterfactuals to be true. On the regular Lewis account, a might-counterfactual $P \Diamond\!\!\rightarrow R$ is true iff there is at least one PR-world among the closest P-worlds. But on the probability view, these can't be the truth conditions. A might-counterfactual is true iff there are *enough* PR-worlds among the closest P-worlds—whatever number is enough to cross the threshold from $P \Box\!\!\rightarrow \neg R$ to $\neg(P \Box\!\!\rightarrow \neg R)$. On this view, the mights and woulds in the cases that motivate counterfactual skepticism would contradict each other *if they were both true* (since mights and woulds are duals, $P \Box\!\!\rightarrow Q$ cannot be true at the same time as $P \Diamond\!\!\rightarrow \neg Q$). But given the cases, the relevant might-counterfactuals *don't come out true*. By hypothesis, the probability theory says that (for example) 2 is true because a large enough proportion of vase-dropping-worlds are vase-breaking-worlds. Thus there's not a large enough proportion of vase-dropping worlds that are not-breaking-worlds to make the might-counterfactual true. In this scenario, the probability theorist faces two problems: how to explain the intuitive truth of the might-counterfactual, and how to explain the clash. If the probability theorist gives up on the duality of woulds and mights, this could open the door for explaining why they are both true, but not why they seem to clash with each other. Nor can the probability theorist explain the retraction data—the relevant would-counterfactuals stay true across contexts, on this view.

Of course, the probability theory could solve some of these problems by adding some context-sensitivity along the lines I have suggested, where the context-sensitivity determines a changing probability threshold. Such a view is not so different from my own, but it still does worse than the view I've been defending. First, it's not clear that probability does all the work. On the context-sensitive probability view, one would have to stipulate an exception for the actual world (as Bennett 2003 points out); if I say *If you had dropped the vase just now, it would have broken* and in fact, unbeknownst to me you did drop the vase and didn't break due to a quantum event, intuitively I have said something false. In addition, low probability is not the only measure for legitimately ignoring worlds. For example, in the parade case, there are certain contexts in which we can legitimately ignore worlds in which Sophie gets stuck behind someone tall, but not worlds in which she's in the bathroom during Pedro's dance, even if these worlds are of equal probability. Better to incorporate probability into the relevance view (as I did) than incorporate context-sensitivity of threshold into the probability view.

If that wasn't reason enough, the probability view also famously encounters the agglomeration problem. Agglomeration is an overwhelmingly intuitive principle:

Agglomeration: $\mathrm{P} \mathrel{\Box\!\!\rightarrow} \mathrm{Q},\ \mathrm{P} \mathrel{\Box\!\!\rightarrow} \mathrm{R} \vdash \mathrm{P} \mathrel{\Box\!\!\rightarrow} (\mathrm{Q} \text{ and } \mathrm{R})$

As Hawthorne points out, agglomeration fails for the simple reason that for any high probability threshold *n* (below 1), the number of Q-worlds can cross the threshold and the number of R-worlds can cross the threshold without the number of QR-worlds crossing the threshold. Worse still, we can completely divide any outcome (such as the vase breaking) into low-probability sub-cases that are below the threshold so that the theory predicts that for each sub-case Q_n, if $\mathrm{P} \mathrel{\Box\!\!\rightarrow} \neg\mathrm{Q}_n$. By agglomeration it follows that if P, *none* of the subcases would occur, and this is clearly wrong, since by hypothesis they were exhaustive! By contrast, my view preserves agglomeration, as long as the context has not shifted.[17] On the contextualist account, there are no cases of pairs of counterfactuals (or small groups of several counterfactuals) in which each individual counterfactual is true (because the consequents lie just over the probability threshold) while the agglomerated counterfactual is false (because the conjunction of the consequents lie just under the probability threshold). This is because though I take low vs. high probability into account on my view, there are no sharp probability cutoffs, even within a conversation. Of course, if one considers *enough* counterfactuals with the same antecedent and different consequents, eventually low probability events (that were perhaps legitimate to ignore at the beginning of the conversation) could add up to a high probability event that one cannot legitimately ignore. But in this case, either the context has shifted (by forcing the consideration of more and more possible outcomes of a single antecedent in a string of counterfactuals) or the counterfactuals in question were not true from the beginning of the conversation, given its focus from the outset on low probability outcomes. (For example, in the case in which one divides the outcomes of the vase breaking into low-probability sub-cases—say, micro-physical descriptions of the ways in which the vase could fall—one is already in a situation in which the standards of precision are high or low probability outcomes are salient, and so the relevant counterfactuals come out as false, and can't be used as premises in agglomeration.)

5.3. Epistemic Theories

Epistemic accounts, like those espoused by Stalnaker (1968) and DeRose (1999), explain the clash between woulds and mights (in fact, the clash is what motivates DeRose's position). On this view, might-counterfactuals are epistemic possibility operators that take wide scope over would-counterfactuals, that is:

$P \Diamond\!\!\rightarrow R$ is true iff $\Diamond_e(P \Box\!\!\rightarrow R)$

On this theory, woulds and mights don't contradict each other in the clash cases, since the truth of $P \Box\!\!\rightarrow Q$ and $\Diamond_e(P \Box\!\!\rightarrow \neg Q)$ are compatible. Rather they pragmatically clash. The truth of the might-counterfactual makes the relevant would-counterfactual *unassertable* in the same context; this is why we feel as though they contradict each other.

I have less to say against the epistemic account, but it does encounter two significant problems. First, the approach doesn't explain why any would-counterfactuals are true in the first place, given traditional similarity orderings.[18] Second, it is committed to the thesis that all might-counterfactuals have epistemic force, which is contrary to many people's intuitions (unlike DeRose, even Stalnaker thinks some might-counterfactuals have metaphysical force).

To account for why any would-counterfactuals are true in the first place, the epistemic account would have to take on board my suggestion that relevance matters in addition to similarity. Then much of the story I've told would go similarly on the epistemic account. Instead of actually shifting the worlds in the domain, might-counterfactuals shift the standards of precision for knowledge, making salient the possibility that we were wrong about which worlds are relevant. I'm not entirely opposed to this move, though I think it is unlikely that unembedded mights have a range of modal force and mights embedded in counterfactual conditionals only have epistemic force.

5.4. Error Theories

The final option is to embrace counterfactual skepticism, as Hájek does. In this case, one needs to adopt an error theory to explain why we go around using counterfactuals that express falsehoods. Hájek himself wants to tell such a story: we use counterfactuals in a sloppy way, because they express something close enough to the truth. On the error theory view, when we say things like 2, they are false, but there is a true counterfactual in the vicinity—one with an explicit probability in the consequent or a specified enough antecedent. We are guilty of rounding errors—treating highly probable consequents as certain. Why not embrace counterfactual skepticism when there's this error theory to explain our communicative behaviour so readily available?

I would find the error theory a lot more convincing if we were just dealing with the esoteric examples from physics. After all, our communicative habits developed before anyone knew anything about quantum mechanics or statistical mechanics. As Hájek says in the introduction to his manuscript, the project of making sense of ordinary utterances and the project of making sense of the universe may not line up, since ordinary speakers may simply be ignorant of the true nature of the universe. I wholeheartedly agree. But, as I've argued, counterfactual skepticism is not singularly a result of advanced scientific understanding of the universe that post-dates the evolution of counterfactual talk. It is a very general feature of our use of counterfactuals that we ignore ordinary, not even terribly improbable possibilities, so our habits can't be explained by scientific ignorance. In doing semantics of natural language, I aim to account for the meaning of counterfactuals in the mouths of native speakers. This is to say, I aim to explain our communicative habits in general, particularly when we are not mistaken about any empirical matters. (I do not think we should aim to explain every specific intuition regarding particular counterfactuals, as some of our intuitions may simply be wrong, and it is certainly not the task of a semantic theory to account for each one.) To say that competent, native speakers systematically mis-use the counterfactual conditional should be our last resort, only if no good semantic story can be told. And I have argued that a good one can indeed be told.

6. A Remaining Challenge: Counterfactuals in Theoretical Contexts

I've been arguing that we should accept a contextualist semantics for counterfactuals, one that takes into account not only similarity in the ordering of worlds, but also relevance. This saves our ordinary counterfactuals—ordinary counterfactuals in the mouths of people in every day conversation can and do often come out as true. But I've said nothing yet about the theoretical role of counterfactuals. Counterfactuals play important roles in decision making, machine learning, history, and psychology. Moreover, many philosophers want counterfactuals to do heavy duty work in explaining causation, laws of nature, free will, dispositions, and more. I contend that counterfactuals can be true in many of these domains, but what counts as a relevant possibility has to take into consideration not only the particular conversation, but what is relevant to the domain of inquiry in general. (Or in other words, a conversation that takes place within a particular domain of inquiry must always take into consideration the possibilities relevant to that domain.)

One might worry that counterfactuals, if they are context-sensitive in the way I have argued, are not really cut out for work in scientific inquiries and domains related to scientific inquiries (such as philosophy of science or causation). These contexts are natural candidates for ones in which low probability outcomes are always relevant. Thus it is possible that we are left with a limited counterfactual skepticism: most counterfactuals in scientific theoretical contexts are false. However, I have some optimism that this is not the case. While it is true that (non-probabilistic, contingent) counterfactuals about fundamental physics will be false, many people think the special sciences are multiply realizable. That is, truths within the science are stable across a range of possible micro-physical realizations. If this is true, then it might be a good argument for the case that the possibilities raised by the fundamental physics—quantum physics or statistical mechanics—aren't relevant to the special sciences (or the philosophy of these sciences), and so counterfactuals can be true in these contexts after all. I leave this as a question for future research.

Counterfactuals, on the view I've been espousing, turn out to be elusive in a similar sense to that in which Lewis (1996) described knowledge as such. We think we have a hold on them—after all we use them effectively all the time—but examine them, and they seem to slip through our fingers. Contextualism, I've argued, makes sense of use of counterfactuals in conversation, both vindicating our judgments of those that are true, and explaining the inescapable clashes between woulds and mights.

Notes

[1] See Lewis 1973 and Stalnaker 1968.

[2] This is how Dodd 2011 explicitly sets up the problem, but it is implicit in much other work on the subject, since the various other proposed solutions—discussed in §5—all involve rejecting at least one of these.

[3] Giving this up is counterfactual skepticism.

[4] Hawthorne 2005 makes a similar argument regarding the quantum cases and Bennett 2003 make a similar one regarding both the quantum and statistical mechanics cases.

[5] The limit assumption, which Stalnaker holds and Lewis rejects, is that there is a set of closest worlds (as opposed to worlds getting infinitely closer and closer without ever reaching a sphere we can call closest). Whether we accept or reject the limit assumption is irrelevant to the present paper, and so I will accept it because it makes the truth conditions of counterfactuals a lot easier to talk about. (I also happen to believe that it should be adopted.) The uniqueness assumption, which Stalnaker also holds and Lewis also rejects, is that there is a unique closest world in every context. The uniqueness assumption is somewhat relevant to the topic at hand. If there is really a unique closest world in every context, then there is no problem of counterfactual skepticism as I am currently motivating it, for there cannot be any problematic worlds *among* the closest worlds, since there is just one closest worlds. And might-counterfactuals don't actually contradict woulds, because they can't be their duals. I will address this in §5.3 below.

[6] Lewis himself does not put the similarity ordering quite this way, because he avoids using temporal locutions like "before the time of the fork", since he wanted his theory of counterfactuals to account for the direction of time. Since I'm not concerned with that here, I use the simpler presentation of the same idea. The reader who prefers Lewis's formulation is free to subsitute it in for this one, without any effect on my argument.

[7] A notable exception is Jonathan Ichikawa 2011, who argues for a contextualist thesis is the same spirit as mine, drawing on a close analogy with contextualism for knowledge. Among the main differences between my view and Ichikawa's include the fact that he motivates counterfactual skepticism through data from reverse Sobel sequences, not the clashes between woulds and mights nor the problem of undermining possibilities among the closest worlds, and as a result his contextualist theory is not concerned with novel ordering sources or with might-counterfactuals, as mine is.

[8] I am employing the comparison to absolute gradable adjectives only as a heuristic. I do not mean to make the much stronger, and presumably false, claim that counterfactuals, or some part of them, have the same semantics as an absolute gradable adjective.

[9] See for example the following papers: von Fintel & Iatridou 2008, Katz et al. 2012, and Charlow 2013.

[10] Also, for Lewis, a quasi-miracle makes a world less similar to the actual world in a permanent way, so it's not a contextualist solution to the problem at all; my point is just that my notion of contextualism is *not* a contextualist version of Lewis's notion of a quasi-miracle.

[11] Similar considerations apply to the fact that Sophie has normal bladder function, the parade had bathrooms, and so Sophie might have been in the bathroom during Pedro's dance. Similarly, Sophie might have gotten an important phone call, etc.

[12] Gillies argues that $\Diamond(P\&R)$ is a presupposition of $P \Diamond\!\!\rightarrow R$ and that the semantics of counterfactuals need to be treated dynamically because of this. I don't agree that mights have this sort of presupposition nor do I agree with his specific treatment of them, but I certainly agree with the central insight behind it.

[13] This contrasts with Gillies' position, wherein might-counterfactuals can expand the domain very far (where the upper bound is perhaps just the laws).

[14] DeRose 2009 takes a similar strategy in showing that these sorts of dialogues are often infelicitous when it come to the gradable adjective *tall*.

[15] DeRose 2009 also also mentions in fn. 13 on p.173 that he does not think the propositional anaphora in Stanley's is good.

[16] DeRose 2008, 157 makes a similar observation, though to argue for a different conclusion.

[17] This caveat about change in context was already the case for the traditional Lewis-Stalnaker theory, given that one does not want to validate a move from (20) and (21) to *If Caesar had been in command in Korea, he would have used catapults and the atom bomb*.

[18] Recall that on Stalnaker's view, the contextually determined selection function picks out a unique closest world. Stalnaker himself thinks that many contexts fail to pick out a unique world, and defends a supervaluationist account of what goes on in such cases. Unless we include relevance in the selection function for choosing a unique closest world, we'd be left with a version of counterfactual skepticism that says most counterfactuals are indeterminate.

References

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

Charlow, Nate. 2013. What We Know and What to Do. *Synthese* 190(12). 2291–2323.

DeRose, Keith. 1999. Can It Be That It Would Have Been Even Though It Might Not Have Been? *Noûs, Supplement: Philosophical Persepctives* 33(13). 385–413.

DeRose, Keith. 2008. Gradable Adjectives: A Defence of Pluralism. *Australasian Journal of Philosophy* 86(1). 141–160.

DeRose, Keith. 2009. *The Case for Contextualism*. Oxford University Press.

Dodd, Dylan. 2011. Quasi-miracles, Typicality, and Counterfactuals. *Synthese* 179. 351–360.

Elga, Adam. 2004. Infinitesimal Chances and Laws of Nature. *Australasian Journal of Philosophy* 82. 67–76.

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

von Fintel, Kai & Sabine Iatridou. 2008. How to Say *Ought* in Foreign: The Composition of Weak Necessity Modals. In J. Guéron & J. Lecarme (eds.), *Time and Modality*, Springer.

Gaifman, Haim & Marc Snir. 1982. Probabilities over Rich Languages, Testing and Randomness. *The Journal of Symbolic Logic* 47(3). 495–548.

Gillies, Thony. 2007. Counterfactual Scorekeeping. *Linguistics and Philosophy* 30. 329–360.

Hájek, Alan. ms. Most Counterfactuals are False. ANU, monograph in progress.

Hawthorne, John. 2004. *Knowledge and Lotteries*. Oxford University Press.

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

Ichikawa, Jonathan. 2011. Quantifiers, Knowledge, and Counterfactuals. *Philosophy and Phenomenological Research* 82(2). 287–313.

Katz, Graham, Paul Portner & Aynat Rubinstein. 2012. Ordering Combination for Modal Comparison. In *Proceedings of SALT* 22, 488–507.

Lewis, David. 1973. *Counterfactuals*. Blackwell.

Lewis, David. 1979. Scorekeeping in a Language Game. *Journal of Philosophical Logic* 8. 339–359.

Lewis, David. 1986. Counterfactual Dependence and Time's Arrow. In *Philosophical Papers Volume II*, chap. 17. Oxford University Press.

Lewis, David. 1996. Elusive knowledge. *Australasian Journal of Philosophy* 74(4).

Stalnaker, Robert. 1968. A Theory of Conditionals. In N. Rescher (ed.), *Studies in Logical Theory*. Oxford University Press.

Stalnaker, Robert. 1981. A Defense of Conditional Excluded Middle. In Harper et al. (eds.), *Ifs: Conditionals, Belief, Decision, Chance, and Time*, Dordrecht: D. Reidel.

Stanley, Jason. 2005. *Knowledge and Practical Interests*. Oxford University Press.

Unger, Peter. 1975. *Ignorance: A Case for Skepticism*. Oxford: Clarendon Press.

Williams, Robert G. 2008. Chances, Counterfactuals, and Similarity. *Philosophy and Phenomenological Research* 77(2). 385–420.