Heim Sequences and Why Most Unqualified 'Would'-Counterfactuals are Not True

Yael Loewenstein

Published in *Australasian Journal of Philosophy* (2020)

If citing, please refer to the published version at
https://www.tandfonline.com/doi/full/10.1080/00048402.2020.1792954

Abstract

The apparent consistency of *Sobel sequences* (example below) famously motivated David Lewis to defend a variably strict conditional semantics for counterfactuals.

(a) If Sophie had gone to the parade she would have seen Pedro.
(b) If Sophie had gone to the parade and been stuck behind someone tall she would not have seen Pedro.

But if the order of the counterfactuals in a Sobel sequence is reversed – in the example, if (b) is asserted *prior* to (a) – the second counterfactual asserted no longer rings true. This is the *Heim sequence* problem. That the order of assertion makes this difference is surprising on the variably strict account. Some argue that this is reason to reject the Lewis-Stalnaker semantics outright. Others argue that the problem motivates a contextualist rendering of counterfactuals. Still others maintain that the explanation for the phenomenon is merely pragmatic. I argue that none of these are right, and defend a novel way to understand the phenomenon. My proposal avoids the problems faced by the alternative analyses and enjoys independent support. There is, however, a difficulty for my view: it entails that many ordinarily-accepted counterfactuals are not true. I argue that this (apparent) cost is acceptable.

Keywords: Counterfactuals, Semantics of Counterfactuals, Heim Sequences, Reverse Sobel Sequences, Counterfactual Contextualism, Variably Strict Conditional Analysis

1. Introduction

Sophie is considering going to a parade where baseball player Pedro Martinez will be featured on a float. She decides not to go. The following counterfactual seems true:

(1) a. If Sophie had gone to the parade she would have seen Pedro.

Now consider what would have happened if Sophie had gone to the parade but had been stuck behind someone tall. It seems that (1b), below, is true as well:

b. If Sophie had gone to the parade and been stuck behind someone tall she would not have seen Pedro.

(1) is an example of what is known as a *Sobel sequence*.[1]

Sobel Sequence Schematic
$p \;\Box\!\!\rightarrow q$, [but] $((p \;\&\; r) \;\Box\!\!\rightarrow \neg q)$

That both counterfactuals of a Sobel sequence can seemingly be true in a fixed context famously motivated David Lewis [1973] to reject the once popular strict conditional semantics for counterfactuals – which treats a counterfactual as a material conditional embedded under a necessity operator – in favour of the now classic *variably* strict semantics. The latter has accessible worlds ordered according to their similarity to the world of assessment, *w*, based on a particular, contextually-determined comparative similarity relation. If '>' represents the counterfactual connective, then for Lewis (but using my notation) a counterfactual A>C is non-vacuously true just in case there is an accessible A&C-world closer to *w* than any A& Not-C world. It is vacuously true just in case there are no accessible A-worlds, and it is false otherwise.[2]

The variably strict semantics has no trouble predicting that both (1a) and (1b) are non-vacuously true in a fixed context. If all the Sophie-goes-to-the-parade-worlds most similar to *w* are worlds at which Sophie sees Pedro, then (1a) is true. If all the Sophie-goes-to-the-parade-*and*-gets-stuck-behind-someone-tall-worlds most similar to *w* are worlds at which Sophie does not see Pedro, then (1b) is true as well. As long as the most similar

[1] Sequences of this form are first discussed in Sobel [1970]

[2] Stalnaker's [1968] semantics is very similar. For Stalnaker 'A>C' is (nonvacuously) true just in case C is true at *the closest* A-world.

worlds at which Sophie goes to the parade and gets stuck behind someone tall are less similar to *w* than the most similar worlds at which she goes to the parade, both counterfactuals can be non-vacuously true.

The problem of Heim sequences[3], attributed to Irene Heim and first discussed by Kai von Fintel [2001] and Anthony Gillies [2007], is that if the order of the counterfactuals in a Sobel sequence is reversed, it no longer seems both counterfactuals are true. Consider:

(2) a. If Sophie had gone to the parade and been stuck behind someone tall, she would not have seen Pedro.

b. #But if Sophie had gone to the parade she would have seen Pedro.[4]

The variably strict semantics seemingly fails to predict the infelicity of (2b). If the closest worlds at which Sophie goes to the parade are worlds at which she sees Pedro (as (1a)/(2b) says), and if the closest worlds at which she goes to the parade and is stuck behind someone tall are worlds at which she does not see Pedro (as (1b)/(2a) says), then the account predicts that the counterfactuals in sequence (2) should both be true as well, regardless of which is uttered first. The analysis appears unable to explain why the order of assertion makes a difference to the counterfactuals' truth-values.

This challenge for the classic model is taken seriously in the literature. It has led some theorists (von Fintel [2001], Gillies [2007]) to reject the variably strict semantics entirely, and instead endorse a variation of the original strict conditional account. Others (Jonathan Ichikawa [2011], Karen Lewis [2018]) argue that the problem motivates a contextualist rendering of counterfactuals similar to contextualist accounts of knowledge or taste. And an entirely different kind of response comes from Sarah Moss [2008], who

[3] Some previous authors have referred to sequences like (2) as *reverse Sobel sequences.* I follow Karen Lewis [2018] in referring to them as *Heim sequences.*

[4] The "#" sign is used to indicate the infelicity of the utterance.

argues that in fact, the classic semantics can handle Heim sequences like (2) just fine. On her view the infelicity of (2b) has a pragmatic explanation and should not be attributed to the counterfactual being *false*.

I contend that none of these are quite right. After showing why I think each is wrong, I will defend a novel way to make sense of the sequences which is also compatible with the classic semantics. My solution avoids the problems faced by the alternative analyses. Additionally, there is good independent reason to think it is right. There is, however, a difficulty for my view: it entails that many ordinarily accepted counterfactuals are *not* true.[5] I will argue that this (apparent) cost is an acceptable one. But first, we should take a brief look at the extant proposals.

2. Von Fintel and Gillies: The Dynamic Semantic Solution

In response to the Heim sequence problem, von Fintel [2001] and Gillies [2007] have (independently) argued that Lewis and Stalnaker were wrong to reject a strict conditional analysis in favour of a variably strict semantics, in the first place. Instead, they contend, the strict conditional semantics is basically correct – it just needs some tweaking. Recall that the strict conditional analysis treats a counterfactual as a material conditional embedded under a necessity operator; the *accessible* worlds are the worlds in the necessity operator's domain (in the context). For example, (1a) is true just in case all of the accessible worlds at which Sophie goes to the parade are worlds at which she sees Pedro.

Von Fintel and Gillies defend a variation of the strict conditional semantics according to which, as part of its *meaning*, a counterfactual utterance 'A>C' updates the domain of worlds that it, and subsequent counterfactuals, quantify over: in particular, 'A>C'

[5] I say *not true* rather than *false* to leave open the possibility that counterfactuals can have indeterminate truth values.

demands that there are accessible A-worlds in the domain.[6] These *dynamic semantic* analyses account for the infelicity of (2b) as follows. As part of its meaning, (2a) demands that there are at least some accessible worlds at which Sophie goes to the parade and gets stuck behind someone tall. But then (2b) is false: it is not the case that Sophie sees Pedro at all accessible worlds at which she goes to the parade. However, although (2b) is false, (1a) need not be. Because (1a) is asserted prior to (1b), at the time of (1a)'s utterance there has been no demand that there are accessible worlds at which Sophie goes to the parade and gets stuck behind someone tall. And (1b) need not be false, either. (1a) demands only that there are some accessible worlds at which Sophie goes to the parade. Accommodating the (weaker) demands of (1b) requires that we bring worlds at which Sophie gets stuck behind someone tall into the domain, and at these worlds, Sophie does not see Pedro. Thus, (2a) and (2b) are semantically inconsistent, but (1a) and (1b) are not. This accounts for why sequence (1) is felicitous while (2) is not. As we will see in the next section, however, there are good reasons to reject the dynamic semantic account.

3. Moss: the Pragmatic Solution

Moss (2012) defends an entirely pragmatic way to account for the infelicity of Heim sequences. On Moss's view, (2b) is infelicitous for the same basic reason that, in general, when an assertion is made in a context in which there is a salient possibility that (i) cannot be ruled out by the speaker and (ii) is incompatible with the assertion, the assertion is infelicitous.[7] Since the speaker cannot rule out that if Sophie had gone to the parade she

[6] The details regarding exactly how this occurs (and which distinguish von Fintel's account from Gillies's) are not important for my purposes here.

[7] Pragmatic infelicity of this kind is a widely occurring phenomenon instantiated by a variety of different kinds of utterances – not just counterfactual ones. Moss discusses some examples in her [2012].

would have been stuck behind someone tall, the possibility, when raised to salience by (2a), makes (2b) pragmatically infelicitous.[8]

Moss gives two considerations that rule in favour of her pragmatic explanation over von Fintel's and Gillies' semantic one. First, the pragmatic explanation accounts for a wider range of data. It explains intuitions not only about Heim sequences but also about infelicitous counterfactuals in other linguistic contexts. For example, assuming that Sophie getting stuck behind someone tall cannot be ruled out by the speaker, (2b) will be infelicitous if asserted after (3), even though (3) is not a counterfactual.

> (3) Remember when Kate got stuck behind a tall person and missed seeing Pedro in her first baseball parade? (2012: 578)

While the pragmatic account is general enough to accommodate examples like these, the semantic solution is not: since (3) is not a counterfactual, it does not as part of its meaning expand the domain over which (2b) quantifies; so there is no prediction that (2b) will be infelicitous if it follows (3).

The second consideration is that Moss's proposal provides a natural explanation for why in some instances the second counterfactual of a Heim sequence does *not* sound infelicitous. As we will see in §4 and §5, there are many examples of sequences that are like this. If the speaker can rule out the possibility made salient by the first counterfactual, then Moss's first criterion for pragmatic infelicity is not met and infelicity is no longer predicted. On the other hand, felicitous Heim sequences *are* a problem for von Fintel and Gillies. If it is part of the meaning of the first counterfactual of a Heim sequence that it expands the

[8] (1a) does not similarly make (1b) infelicitous because (1a) is intuitively no longer part of the common ground once (1b) has been uttered. See Moss [2012].

domain over which the second quantifies, then the second counterfactual should be false and infelicitous.

Moss's solution is appealing. There are, however, some problems with it. Lewis (2018) recently advanced a series of objections to Moss's account. In §V I will offer a novel argument which aims to show that, at a minimum, Moss's account cannot tell the full story. First, I consider one last proposal.

4. Karen Lewis: the Contextualist Solution

Karen Lewis ([2016], [2018]) defends an alternative proposal that evades some of the problems faced by previous accounts. Lewis argues that counterfactuals should be understood as on a par with context-sensitive expressions like, for instance, automatic indexicals (*I*, *today*), gradable adjectives (*tall*, *rich*) and absolute gradable adjectives (*flat*). On her view the semantic value of a counterfactual is sensitive to features of the conversational context such as the standards of precision and the salience (or non-salience) of possibilities. Formally, the difference between Lewis's semantics and the Lewis-Stalnaker semantics is that while on the classic model possible worlds are ordered only according to how similar they are to the world of assessment, for Karen Lewis worlds are ordered by both similarity *and relevance*.

> [B]egin with a Lewisian-like similarity metric…[r]elevance takes worlds that are among the most similar worlds and moves them farther away, so that they are not among the closest worlds. It also takes worlds that are less similar – as long as they are similar enough (which is vague) – and moves them closer, so that they are among the closest worlds. [2018: 500, Lewis's emphasis]

Her truth conditions are given below.

> For all contexts c, P $\Box\!\!\rightarrow$ Q is true at *w* in c iff all the closest P-worlds to *w* are Q-worlds, where closeness is a function of both similarity and relevance. [2018: 500]

Lewis maintains that not every possibility can be made relevant if brought to salience and not every possibility can be made irrelevant by being ignored: there are objective constraints on what must, can, and cannot be relevant in a context. For example, "High probability (conditional on the antecedent) macroscopically-described outcomes are always relevant" and "…some possibilities are just determinately irrelevant, like ones that are really dissimilar to the actual world…" [2018: 500–501]

How does this solve the Heim sequence problem? Return to sequences (1) and (2). Supposing that (1a), (1b) and (2a) are true, (2b) can still be false if the possibility raised to salience by (2a) – that is, Sophie is both at the parade and stuck behind someone tall – is relevant in the context when (2b) is evaluated. If it is relevant then worlds at which Sophie goes to the parade and is stuck behind someone tall become among the closest Sophie-goes-to-the-parade-worlds (even though they were not among the closest when (1a) was evaluated, which is why (1a) could be true). And if worlds at which Sophie is stuck behind someone tall and so does not see Pedro are among the closest Sophie-goes-to-the-parade-worlds, (2b) is false.

Lewis's analysis avoids many of the difficulties faced by the alternative proposals.[9] For example, it can account for felicitous Heim sequences: if the possibility raised to salience by the first counterfactual of the sequence is not relevant, it does not impact the closeness ordering. Nevertheless, I think the contextualist solution is wrong. I now advance two arguments against it.

[9] Lewis discusses some advantages of her account over Moss's in her [2018].

Argument 1: Counterfactuals are not like the others.

One immediate challenge for a contextualist semantics like Lewis's is that counterfactuals like (1a) seem to have invariant truth values once the "facts have been fixed" and the similarity ordering has been established.[10] It is clear that the truth-value of a statement like <Agnes is rich> or <Lester is tall> depends on the standards for richness, or for tallness, in the context. But counterfactuals seem importantly different. Intuitively, it is *not* the case that whether Sophie would have seen Pedro had she gone to the parade depends on standards of any kind. Indeed, once the similarity ordering has been established, whether Sophie would have seen Pedro seems to have nothing whatsoever to do with what occurs at the context of utterance. And if that is right, Lewis's contextualism is wrong.

This difference between counterfactuals and paradigmatic context-sensitive expressions might be made more apparent by comparing discourses involving both. When we say a pool table is flat, or a fifth-grade child is tall, we are in general saying something about the table, or the child, relative to some standard or comparative class. If it is objectively true that the table is flat (or the child tall), it is so relative to the standard implicit in the conversational context. Indeed, as we will see in a moment, the speaker is usually able to make explicit the standards or comparative class or domain she has in mind when pressed.

But a counterfactual utterance is different. When I assert, <if Sophie had gone to the parade she would have seen Pedro>, I do not take myself to be asserting what would have happened relative to a standard of any kind. The difference is evident upon consideration of some examples:

[10] Of course, most agree that counterfactuals are context-sensitive in one sense: which facts count toward the world similarity ordering depends upon the context. Lewis distinguishes her own kind of context-sensitivity from this "ordinary" kind by saying that hers "is context-sensitivity after all the facts are fixed." [2016: 16]. In other words, there is additional work for the context to do once the similarity function is (contextually) determined.

(4) a. He is tall!

b. [#?] No he is not. Remember the NBA players we saw at the game?

c. $_{\text{Ok}}$ Don't be a smart alec. I meant he is tall for a boy his age.

(5) a. The table is flat

b. [#?] No it is not; if you look with a microscope you'll see unevenness…

c. $_{\text{Ok}}$ What I meant was for our purposes it is flat. It is flat enough to lay your drink on.

(6) a. That woman is wealthy.

b. No she is not. Have you been to my neighbourhood?

c. $_{\text{Ok}}$ Relative to my standards she is wealthy. Relative to your standards she is not.

Compare:

(7) a. If Sophie had gone to the parade she would have seen Pedro.

b. $_{\text{Ok}}$ But if Sophie had gone to the parade and been stuck behind someone tall she wouldn't have seen Pedro.

c. # That's not what I meant. I meant that for our purposes she would have seen Pedro./# Okay, relative to my standards she would have seen Pedro but relative to your standards she might not have./#That's not what I meant, I meant that if Sophie had gone to the parade she would have seen Pedro relative to standard (or restricted by domain) ___. [Where in the "___" goes any standard (or domain).]

The problematic replies in (7c) provide good evidence the speaker who asserts (7a) takes herself to be speaking in absolute terms. She does not intend to assert that <if Sophie had gone to the parade she would have seen Pedro> is true relative to some standard or other.

While most I have informally surveyed agree with me that the variations of (7c) are infelicitous, a few have the intuition that in ordinary contexts the intended meaning of (7a) is in fact something like, (7a') <*assuming things would have gone as expected*, or, *given ordinary circumstances*, if Sophie had gone to the parade she would have seen Pedro>. It would be futile to try to deny that (7a) could ever be used to mean this. We do have reason to deny that this is the ordinary or usual meaning, however.

Suppose that unbeknownst to the speaker of (1a), Sophie actually went to the parade, was stuck behind someone tall and did not see Pedro. Would we then think the speaker spoke falsely? It seems clear we would; and this regardless of the reason Sophie did not see Pedro. But that is difficult to explain if the current proposal is right. If (1a) means <if Sophie had gone to the parade *and nothing unexpected happened* then she would have seen Pedro>, then its truth is *compatible* with Sophie actually going to the parade, getting stuck behind someone tall and not seeing Pedro. And if that is so, we should not judge the utterance as clearly false when it turns out she actually was stuck behind someone tall. That we judge it false is good evidence that (7a') is *not* (1a)'s meaning in ordinary contexts.

Can weak centering – the principle that the world of assessment is always among the worlds closest to itself (which entails that if the antecedent is true and consequent false the counterfactual is not true) – account for why we would evaluate (1a) as false if it turns out that the antecedent is true and the consequent false? It cannot, given the proposal presently considered. According to that proposal, the meaning of (1a) is something like (7a') <if Sophie had gone to the parade *and nothing unexpected happened* she would have seen Pedro>. Since in our imagined scenario it turns out that Sophie actually went to the parade and was stuck behind someone tall, the antecedent of (7a') is *false*. If the antecedent is false then centering is of no help, and my objection to the proposal that (1a) really means something like (7a') still stands.

Perhaps contextualism about counterfactuals should be seen as more akin to contextualism about knowledge.[11] Lewis declines to compare her account to knowledge contextualism because, as she says, she does not want to rest her defence on controversial cases of alleged context sensitivity. [2015:18] Nevertheless, since knowledge contextualism has faced similar objections to what I advance here, we should briefly consider whether the distinct contextualist projects are in the same boat. In response to a similar kind of objection aimed at knowledge contextualism, Keith DeRose [2008], [2009] argues that speakers do sometimes clarify what they meant by previous knowledge assertions by making explicit the epistemic standards active at the time of utterance. For instance, when pressed about an earlier knowledge claim made in a context with lower epistemic standards, a speaker might reply: "all I was claiming was that I know it by ordinary standards/beyond a reasonable doubt…I never meant to be claiming to know it for certain/beyond all possible doubt." [2008: 143]

This style of reply is not readily available to the contextualist about counterfactuals. If knowledge contextualism is right, we might expect speakers to be able to clarify previous knowledge assertions by making explicit the epistemic standards active at the time of utterance, since it is these standards that allegedly vary. And if DeRose is right, this is exactly what speakers sometimes do. Similarly, if Lewis's contextualism is right, speakers should be able to make explicit the standards of relevance active at the time of utterance, since it is these that Lewis takes to vary. But this is precisely what I have argued should be denied. We saw that making explicit the standards of relevance as in (7c) is clearly infelicitous. And

[11] I thank an anonymous referee for pressing me on this.

understanding (1a) as something like (7a') is not a good option either, since (7a')'s truth is compatible with Sophie actually going to the parade and getting stuck behind someone tall; yet we'd judge (1a) as clearly false if this happened.[12] DeRose's strategy for defusing the objection to knowledge contextualism is not applicable here. While the difficulty in making explicit the standards alleged by Lewis to be implicit in counterfactual assertions does not prove that the standards are not there, it provides good evidence for their absence, especially given that there appears to be no similar trouble for context-sensitive expressions of other sorts.

<u>Argument 2</u>: Lewis's semantics makes the wrong predictions.

Compare our original Heim sequence, sequence (2), to sequence (8), below.

You and I are talking about a particular (dry) match:

(8) a. If this match had been wet and struck at time t it would not have lit.

b. $_{\text{Ok}}$Okay, but if this match had been struck at t it would have lit.

Why is (2b) infelicitous but not (8b)? Karen Lewis's semantics predicts that Heim sequences will be felicitous if the possibility made salient by the antecedent of the first counterfactual in the sequence obtains *only at worlds too far away to be relevant*. But why are the wet-match worlds too far away? Surely we can devise a scenario where the wet-match worlds are not so far away. Imagine there are thousands of wet matches in the room, and the match referred to in (8) is the only dry match. Or suppose that at some time shortly before t

[12] Note that the same line does not work against DeRose's proposal for understanding certain knowledge claims. If a speaker were to find out that the proposition she had previously claimed to know were in fact false, we'd expect her to accept that she had spoken falsely. But here this is not surprising, even if knowledge contextualism is right. If knowledge requires truth, and if the proposition asserted as known is false, then the epistemic standards at the time of utterance make no difference: the knowledge claim will be false.

there is a high probability that the match in question will get wet. Despite the close call, the match remains dry. In both cases it is still felicitous to utter (8b) in response to (8a). Since we are referring to a particular dry match, it appears to make no difference how far away the worlds at which the match is wet are.

Attempting to provide an explanation for why worlds at which the match is wet are too far away to be relevant in ordinary contexts brings out something peculiar: with one exception, no matter how similar we make the worlds at which the match is wet, (8b) remains felicitous. What is the exception? If t stands in for the moment the match first hits the matchbox, let us use t_1 to stand in for the moment, some milliseconds later, when the match catches fire. If there is a chance – whether small or large – that the (actually dry) match could have gotten wet *because of* the process of the striking (and prior to t_{1}), then it would no longer be felicitous to utter (8b) in response to (8a). In this case it is not known that the match would have been dry at t_1 had it been struck at t.

There is one context, then, in which worlds at which the match is wet and struck are not too far away to be made relevant. This is a context where there is a chance that the otherwise dry match could have become wet prior to the flame igniting if the striking had occurred. But notice that in this context we no longer have reason to think that worlds at which the match is struck and wet are further away from the actual world than worlds at which the match is struck and dry. For although the match was actually dry at t_1, this cannot be relevant to the closeness ordering if the process of the match striking could have itself caused the match to become wet. In general, we should not count factors that are causally dependent on the events required for the obtainment of the counterfactual's antecedent toward world similarity. To see this imagine a scenario in which the chance that the match would have become wet if struck is very high, say 1. In that case we certainly can't count the actual dryness of the match at t_1 as relevant for the similarity ordering, since <if this match had been struck at t it would *not* have lit> should come out true. And for it to be true the worlds at which the match is wet and struck must be at least as close as the worlds at which the match is dry and struck (indeed, *closer*, since we'd also want to evaluate <if this match had been struck at t it would have been wet at t> as true).

Unless there is a chance that if the match were struck it might first become wet (perhaps in the process of the striking), it does not matter how similar the wet-match worlds are: it remains the case that (8b) is a felicitous response to (8a). Furthermore, if there *is* a chance that if the match had been struck it could have become wet prior to t_1, we no longer have reason to think wet-match worlds are any further away than dry-match worlds. This suggests that, contra Karen Lewis, *the best explanation for the felicity of (8b) is* ***not*** *that worlds at which the match is wet and struck are vastly dissimilar from the actual world* (since, as we've seen, they need not be very dissimilar at all for (8b) to be felicitous). A much better explanation is that the nearest worlds at which the match is struck and wet are simply *less* similar to the actual world than the nearest worlds at which the match is struck and dry – and only the most similar antecedent worlds are relevant for the evaluation of a counterfactual. In other words, we apply the classic similarity function to rule out worlds that are not among the most similar; and relevance has no work to do.

As far as I can tell, there is nothing special about the match example to warrant thinking that what we've said about this case would not generalise (I encourage the reader to try to come up with a counterexample for herself). And, as we will see in §5, it does seem to generalise: whether a Heim sequence is felicitous or not appears to track whether the possibility made salient by the first counterfactual of the sequence obtains at worlds that are among the *most* similar worlds at which the antecedent of the second counterfactual is true.

How similar or dissimilar the salient worlds makes no difference, as long as they are not among the most similar antecedent-worlds. And if that is right then the similarity function handles the data best on its own.

5. Deflating the Problem?

This suggests a different solution to the problem. Compare our original Heim sequence, sequence (2), with sequence (8). As we've seen, (8b) is felicitous: indeed, one might use (8b) as a way of implying that the match is dry. Why is (8b) felicitous but not (2b)? Here is an answer no one has given: worlds at which Sophie goes to the parade and is stuck behind someone tall are *just as similar* to the actual world as are worlds at which she goes to the parade and sees Pedro. (Of course, it may be *unlikely* that Sophie would have been stuck behind someone tall, but similarity is not a function of likelihood.[13]) In contrast, if the match is actually dry, then assuming that there is no chance that the match would have become wet in the process of the striking (if there were a known chance, (8b) would also be *in*felicitous), worlds at which the match is wet and struck are *less* similar to the actual world than worlds at which the match is dry and struck (and so lights).

There is good reason to think that worlds at which Sophie went to the parade but was stuck behind someone tall are just as close as worlds at which she went to the parade and saw Pedro. To see this, it will be helpful to consider one more example.

[13] There are a handful of philosophers who have argued for probabilistic truth conditions, e.g., a counterfactual is true if and only if the consequent is true at a sufficiently high proportion of the closest antecedent worlds [Bennett 2003] and, a counterfactual is true if and only if the conditional probability of the consequent given the antecedent is sufficiently high [Leitgeb 2012a, 2012b]. Even on these accounts, though, *more likely* does not in general mean *more similar* – it is only once a certain threshold of unlikeliness is reached that the (extremely unlikely) possibilities are ruled out. And there are reasons to reject these sorts of accounts, anyway. One simple reason is that they entail that (sufficiently) likely events are always the events that would have occurred, counterfactually. But this seems wrong. *Sometimes* (often, even) the extremely unlikely happens.

An (ordinarily dressed) child is playing on top of the jungle gym and trips and almost falls. Two observers have the following exchange:

(9) a. It's good that child didn't fall. If she had she would have broken a bone!

b. (#?) Yes, but if the child had been wearing a full-body padded suit and fallen, she would not have broken a bone.

c. $_{\text{ok}}$Uhh, okay….but if that child had fallen, she would have broken a bone./$_{\text{ok}}$Okay, but if *that* child had fallen, she would have broken a bone./$_{\text{ok}}$Okay, but if that child, *just as she is*, had fallen, she would have broken a bone.

This is a clear-cut case. The worlds at which the child is wearing the padded suit and falls are less similar to the actual world than worlds at which she is wearing no padded suit and falls. The various appropriate replies to (9b) (shown in (9c)) are suggestive of a (defeasible) heuristic test which we can dub the *Something in the World Test*.

> **The "Something in the World" Test**: If a world is less similar to the actual world than is another world, then there is something that can (in principle) be "pointed to" in the actual world in virtue of which this is so.[14]

It is possible for the person who asserted (9a) to reply to (9b) by emphasizing that he is not talking about a scenario in which the child is wearing a padded suit. He is talking about *that* child, just as she is now. The speaker can point to something about the world — the child, and what the child is wearing – to make clear that the worlds at which the child is

[14] The "something" to point to need not be a particular object or set of objects (thanks to an anonymous referee for pressing me on this). It can be some fact about kinds of objects in general, or about an actual event or events, or about some regularity (note that if the regularity is not exceptionless, and if this is made salient in the first counterfactual of the Heim sequence, then I'd expect the second counterfactual to be infelicitous – since the world with the exception is just as near to the actual world – in which case the test is not applicable).

wearing a protective suit are not relevant to the conversation. They are less similar, and they are less similar because the child has no protective suit at the actual world.

Sequence (8) passes the test as well. In response to (8b), the first speaker might say "that's true, but if this match had been struck it would have lit", and that would be just fine. But if she wanted she could also choose to emphasize that she is talking about *this* (dry) match. She can do that by emphasizing 'this', or by adding a 'just as is' clause: "that is true, but if this match *as it is now* had been struck, it would have lit". We can identify something in the world — in this case, the dryness of the match — to account for why worlds that are different in that respect are less close. Worlds at which the match is not dry are less similar to the actual world in virtue of the fact that at the actual world the match is dry (unless the match has a chance of becoming wet in the process of the striking, in which case wet-match worlds are no further away, and (8b) is *infelicitous*).

What about sequences (1) and (2)? What can be identified, in the actual world, which plausibly makes it such that worlds at which Sophie is stuck behind someone tall are less similar to the actual world than worlds at which she sees Pedro? I say there is nothing. In reply to (1b), the speaker who asserted (1a) *cannot* helpfully point to Sophie and say "no, I'm talking about *that* Sophie"; "no, I'm talking about Sophie as she actually is" (unless she is tall enough that the possibility that she is stuck behind someone tall can be ruled out – *in which case (2b) would be felicitous*). Nor can the speaker point to anything about the parade: "no, I'm talking about *that* parade (as it actually is)" (unless it is known that attendees stood in single file, *in which case (2b) would be felicitous*). If there is nothing to point to, we conclude by the *Something in the World* Test that the worlds at which Sophie attends the parade but is stuck behind someone tall (and so does not see Pedro) are no further away than the worlds where she attends and sees Pedro. And in that case, (2b) is not true.

I do not think it is a coincidence that the sequences usually used to illustrate the problem of Heim sequences are (1) and (2) (the counterfactuals in these are not among those regularly used as examples in discussions of other counterfactual-related topics). This is not because I think it has been widely recognised that worlds at which Sophie is stuck behind someone tall are no further away than worlds at which she sees Pedro. To the contrary, just about everyone in the literature takes it for granted that (1a) is true: and for (1a) to be true, all nearest Sophie-goes-to-parade-worlds must be Sophie–sees–Pedro–worlds. But it is the very fact that, as it happens, *not* all nearest Sophie–goes–to–parade–worlds are Sophie–sees–Pedro–worlds that accounts for the infelicity of (2b).

My proposal, then, is this: (1a) <If Sophie had gone to the parade she would have seen Pedro> was never true. It was never true because it is not the case that all the closest antecedent worlds are consequent worlds. And if (1a) is not true then (2b) is not true, either. This explains (2b)'s infelicity. But if (1a) is not true, why did it seem true? I suggest that (1a) seemed true simply because the possibility that Sophie could have gone to the parade and been stuck behind someone tall (or been in the bathroom, or been distracted by her phone) did not occur to us. Someone needed to raise the possibility to bring it to our attention that she might not have seen Pedro.

This solution, according to which neither (1a) nor (2b) is true, avoids the problems faced by the alternative proposals. It is general enough to explain why possibilities made salient by non-counterfactual utterances can make counterfactuals like (2b) infelicitous; it can explain why counterfactuals like (8b) and (9c) are felicitous while those like (2b) are not; and it does not commit us to maintaining that counterfactuals are context-sensitive in the way that gradable adjectives are.

For Moss, the infelicity of counterfactuals like (2b) is accounted for on purely pragmatic grounds. For all that Moss's account says, (2b) can be true. But, as I argued a moment ago, (2b) is not true. It is not true because not all closest Sophie-goes-to-parade-worlds are worlds at which Sophie sees Pedro. This is not particular to this example. We've seen in other cases (the match and playground examples, for instance), the felicity or infelicity of the Heim sequence depends on whether the worlds made salient by the first counterfactual in the sequence are among the most similar antecedent-worlds of the second. If they are among the most similar, then the second counterfactual is both infelicitous and not true.

This is not to say that the second counterfactual of a Heim sequence will *never* be infelicitous if true: I do not want to rule out the possibility that a counterfactual can sound infelicitous not because it is not true but because the people in the conversation do not *know* if it is true. It is undoubtedly possible for there to be vague or ambiguous contexts in which it is impossible to discern which worlds count as among the most similar (in that context) and which do not. And in a context like that, I would expect Mossian-type pragmatic considerations to be relevant: if it cannot be ruled out that there are A&not-C-worlds as close as the nearest A&C-worlds, the counterfactual is unassertable, regardless of its truth-value.

Thus, although I wish to leave open the possibility that Moss's proposal can provide the right diagnosis for special cases in which it cannot be determined (for whatever reason) whether the world made salient by the first counterfactual is among the nearest worlds relevant to the evaluation of the second, I reject Moss's account as a general solution to the problem. I have argued that when the second counterfactual of a Heim sequence is infelicitous, this is (generally) because the world made salient by the first counterfactual is as near as the nearest worlds relevant to the evaluation of the second. If that's right, then the second counterfactual is (generally) not true when the sequence is infelicitous. And if that's right, we need not appeal to pragmatic considerations to account for the infelicity in ordinary cases.

6. The Cost of Counterfactual Scepticism

If I am right, philosophers have been wrong to take Heim sequences as a challenge to the classic Lewis-Stalnaker semantics. When used correctly, the classic semantics either rules two identical counterfactuals both true or both not true in a fixed context, regardless of where each occurs in a sequence. But there is a difficulty for my proposal. Its truth entails that many ordinarily accepted counterfactuals are not true. Consider the match scenario, again. We know that worlds at which the match is wet and struck are further away from the actual world than worlds at which the match is dry and struck, and so these worlds are irrelevant to the evaluation of <if the match had been struck it would have lit>. But there may be other worlds, not yet salient, which are just as close to the actual world and at which the match is struck but does not light. For instance, suppose someone mentions the fact that now and then match strikers get unlucky: it is possible to hit the match box at a bad angle so that the match does not light. In that context the counterfactual <if this match had been struck it would have lit> no longer sounds felicitous. And, as I have just argued, if it is infelicitous, this is generally because it is not true. It is not true because there are worlds just as close as the match-is-struck-and-lights-worlds at which the match is struck and does not light. But if that is so then the ordinary counterfactual <if the match had been struck it would have lit> is not true, even when taken by itself. And indeed, nor are most other 'would'-counterfactuals, since as Hawthorne [2005], Hájek [manuscript], Lewis [2016] and others have pointed out, there is usually some world among the closest antecedent worlds at which the consequent does not obtain for some reason or other.

How serious of a cost is this? Many have gone to great lengths to avoid accepting the conclusion that many ordinary counterfactuals are not true.[15] I find the strong aversion to the claim somewhat surprising. Alan Hájek has famously argued that although most ordinary 'would'-counterfactuals of the form 'A>C' are false, there are nearby counterfactuals of the form 'A>probably-C' which are true. I endorse this view. While <if the match had been struck it would have lit> is probably not true, <if the match had been struck it probably would have lit> almost certainly is. Is this really so counterintuitive? One reason for thinking that it is *not* is that imprecision is utterly ubiquitous in ordinary language.[16] To take an ordinary example, suppose that someone is baking bread in addition to cooking some other food. Jasper walks in, smells the bread and asserts "that is the smell of bread!". This gets the point across but is probably not quite right. Bread smells subtly different than what Jasper perceives. What Jasper smells is actually a mixture of bread and a host of other things combined. Examples of this kind are endless. It is not radical to hold that people regularly speak loosely when asserting counterfactuals if people regularly speak loosely in general.

But why should people regularly speak loosely when uttering counterfactuals if to speak a truth one must be more precise? Why not be a bit more careful? For many of the same reasons, I suggest, that Jasper is not more careful when he asserts that what he smells is the smell of bread. For one, Jasper might not in that moment recognise that there are likely other smells adulterating the pure bread smell. And even if he realized that bread does not smell *exactly* like *that*, he would risk misleading his listeners if he said something more precise. If, just to be safe, Jasper said "that is approximately the smell of bread" or "that is a mixture of smells, the most salient of which is bread", he would risk implicating that he knows more

[15] See, for example, Lewis [1986], Bennett [2003], Williams [2008], Ichikawa [2011] and Lewis [2016], among others.
[16] On this see Teller [2011], Braun and Sider [2007], and Elgin [2004].

than he does. For instance, he might implicate that he recognises other smells that are presently diluting the bread smell, or that he is able to distinguish between pure-bread smell and impure-bread smell. Yet, we can suppose, neither of these things is true. The irony is that if he had expressed something more accurate in the first place – for example, that what he smells is approximately the smell of bread – he would have risked implicating falsehoods.

The same is true for the person who asserts a counterfactual. If I were to get into the habit of always adding "probably" before any counterfactual consequent, I would likely inadvertently communicate much that I do not mean. If I were to say, "if that match had been struck it probably would have lit" my listeners might assume that I had some particular reason to think that the match may not have lit. Maybe I know something about the match, or its environment, that they do not. My listeners are likely to assume that I take the odds that the match would have lit had it been struck to be lower than I actually take them to be. To weaken one's counterfactual assertion in this way could suggest some degree of uncertainty that may not actually be felt. For these reasons, and because it can be tiresome and unnecessary for speakers to be precise, and because unlikely, unexpected ways one's utterance could be made false are often not at the forefront of one's mind when one speaks, it is not at all surprising that 'A>C' is regularly said in place of 'A>probably-C', even if the former is merely *close* to something that is true.

7. Conclusion

I have argued that the Heim sequence "problem" is not really a problem at all. Once we are willing to accept that most unqualified counterfactuals are not true, we can recognise infelicitous Heim sequences for what they are: evidence that we've come across a counterfactual that should be qualified with (something like) *probably*. If this is right, there is no need to revise or reject the classic semantics in response to Heim sequences.

Nevertheless, it might prove worthwhile to shift some attention away from trying to understand unqualified 'would'–counterfactuals first and foremost, and toward trying to advance our understanding of the qualified ones.[17][18]

REFERENCES

Barker, Stephen 1999. Counterfactuals, Probabilistic Counterfactuals, and Causation, *Mind* 108: 427-469.

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

Braun, David and Sider, Theodore 2007. Vague, So Untrue, *Nous* 41 (2): 133-156.

DeRose, Keith 2008. Gradable Adjectives: A Defence of Pluralism, *The Australasian Journal of Philosophy* 86: 141-160.

DeRose, Keith 2009. *The Case for Contextualism: Knowledge, Skepticism and Context*, Oxford University Press.

Elgin, Catherine 2004. True Enough, *Philosophical Issues*, 14: 113-121.

Gillies, Anthony 2007. Counterfactual Scorekeeping, *Linguistics and Philosophy*, vol. 30: 329-360.

Hájek, Alan manuscript. *Most Counterfactuals are False*.

Hawthorne, John 2004. *Knowledge and Lotteries*, Oxford: Clarendon Press.

[17] There is still much work to be done if we are to understand the semantics of probabilistic counterfactuals. For an introduction to some of the puzzles regarding how counterfactuals interact with probability operators, see Barker [1999].

[18] For very helpful comments and discussion I am indebted to Arif Ahmed, Juan Comesaña, Terry Horgan, Boris Kment, Hilah Loewenstein, Luis Oliveira, Carolina Sartorio, Matt Schuler, Jason Turner, Alexis Worlock, several anonymous referees, and especially Alan Hájek. I also thank members of Terry Horgan's 2016/2017 graduate work-in-progress group as well as audiences at the 2017 APA Pacific Division Meeting, the 3rd Belgrade Conference on Conditionals, Texas Tech, UCLA, the University of Arizona and the University of Houston.

Hawthorne, John 2005. Chance and Counterfactuals, *Philosophy and Phenomenological Research*, Vol. LXX, No. 2.

Ichikawa, Jonathan 2011. Quantifiers, Knowledge, and Counterfactuals, *Philosophy and Phenomenological Research*, Vol. LXXXII No. 2.

Leitgeb, Hannes 2012a. A Probabilistic Semantics for Counterfactuals: Part A, *The Review of Symbolic Logic* 5: 1.

Leitgeb, Hannes 2012b. A Probabilistic Semantics for Counterfactuals: Part B, *The Review of Symbolic Logic* 5: 1, 85-121.

Lewis, David 1973. *Counterfactuals*. Basil Blackwell Ltd., Malden, MA.

Lewis, David 1986. Counterfactual Dependence and Time's Arrow, in *Philosophical Papers Volume II*, Oxford University Press: 32-52.

Lewis, Karen 2016. Elusive Counterfactuals, *Nous* 50 (2): 286-313.

Lewis, Karen 2018. Counterfactual Discourse In Context, *Nous*, 52:3, 481-507.

Moss, Sarah 2012. On the Pragmatics of Counterfactuals, *Nous* 46:3, 561-586.

Sobel, J. Howard 1970. Utilitarianisms: Simple and General. *Inquiry* 13, 394–449.

Stalnaker, Robert C. 1968. A Theory of Conditionals, in *Studies in Logical Theory*, ed. Nicholas Rescher, vol. 2 American Philosophical Quarterly Monograph Series, 98 - 112. Oxford Blackwell.

Teller, Paul 2011. Two Models of Truth, *Analysis* 71:3, 465-472.

Von Fintel, Kai 2001. Counterfactuals in a Dynamic Context, in *Ken Hale: A Life in Language*, ed. Michael Kenstowicz, MIT Press, Cambridge.

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