COUNTERFACTUALS WITHOUT CAUSATION, PROBABILISTIC COUNTERFACTUALS AND THE COUNTERFACTUAL ANALYSIS OF CAUSATION

by

Yael Loewenstein

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF PHILOSOPHY

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2017

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Yael Loewenstein, titled Counterfactuals Without Causation, Probabilistic Counterfactuals and the Counterfactual Analysis of Causation and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

____ Date: 7-25-2017
Terry Horgan

____ Date: 7-25-2017
Juan Comesaña

____ Date: 7-25-2017
Carolina Sartorio

____ Date: 7-25-2017
Jason Turner

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

____ Date: 7-25-2017
Dissertation Director: Terry Horgan

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Yael Loewenstein

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to my advisor, Terry Horgan, for the patient guidance and mentorship he provided to me, and for the generousness with which he gave his time and insight. Terry's philosophical brilliance is matched only by his genuinely good nature and down-to-earth humility, and I am truly fortunate to have had the opportunity to work with him.

I would also like to express my sincere gratitude to my committee members, Juan Comesaña and Carolina Sartorio, for being extraordinarily helpful and supportive throughout the duration of my graduate studies and for the general collegiality that each of them offered to me over the years. My work and my overall philosophical ability have been vastly improved thanks to Carolina and Juan. I am grateful to Jason Turner for agreeing to join my committee so shortly after coming to UA, and for his fantastic feedback on my papers.

I'd also like to thank Michael McKenna and Stewart Cohen, both of whom have been wonderfully supportive and encouraging to me from the very start. To Michael I owe a special debt of gratitude for encouraging and guiding me through the writing and submission of my paper, "Why the Direct Argument Does Not Shift the Burden of Proof", which has been so important to my development as a philosopher.

I thank Samer Masri for bringing me joy and his unconditional support during the final and most stressful phases of my PhD and job search.

Matthew Schuler was my rock for most of my graduate school life, and I will forever be indebted to him for his patient support and encouragement, and for the inordinate amount of time he spent listening to me talk about my ideas and proofreading my papers; always providing me with extremely helpful suggestions. I'm not sure that I could have done it without his unwavering belief in me.

I am deeply grateful to my parents and sister for their love, encouragement and support.

DEDICATION

For my beloved grandparents, Rae and Joseph Nemovicher, who instilled in me a love for puzzle solving and from whom I inherited a passion for thinking deeply about philosophical questions.

TABLE OF CONTENTS

ABSTRACT

It is near-consensus among those currently working on the semantics of counterfactuals that the correct treatment of counterfactuals (whatever it is) must invoke causal independence in order to rule a particular set of *seemingly true* counterfactuals – including a famous one called *Morgenbesser's Coin* (MC) – true. But if we must analyze counterfactuals in terms of causation, this rules out giving a reductive account of causation in terms of counterfactuals, and is, as such, a serious blow to the Humean hope of reducing causation to counterfactual dependence.

This dissertation is composed of three self-standing articles. In the first article I argue that counterfactuals like MC are *false* contrary to appearances; as is the thesis that the correct semantics of counterfactuals must appeal to causal independence.

In the second article I argue that there are important, widely-held assumptions about difference-making and its relationship to causation which are false, and which may underlie some of the remaining, most threatening objections to the counterfactual analysis of causation.

In the final article I discuss the puzzle of reverse Sobel sequences – an alleged problem for the classic Lewis-Stalnaker semantics for counterfactuals. I argue that none of the extant approaches to the problem are right, and defend a novel solution to the puzzle. If I am correct, reverse Sobel sequences do not threaten the classic analysis. They do, however, give additional evidence for the thesis, forcefully defended by Alan Hájek, that most non-probabilistic 'would'-counterfactuals are false. This motivates placing a stronger emphasis on trying to understand probabilistic counterfactuals first and foremost.

INTRODUCTION

Suppose that Susan is about to toss a fair, *indeterministic* coin. Lucky has a chance to bet 'heads'. His bet has no causal impact on Susan's toss (we can imagine, for instance, that Lucky and Susan are in different countries, and Lucky is watching the toss about to unfold on live television). Lucky declines to bet heads, Susan tosses the coin, and the coin lands heads. Consider the counterfactual commonly called *Morgenbesser's Coin* (MC): "If Lucky had bet heads he would have won the bet". It is generally taken for granted in the literature that MC is true. The apparent truth of MC is what motivated David Lewis and subsequent similarity theorists to think that the maximization of regions of *imperfect match* (in matters of fact)—and not merely regions of perfect match—is sometimes relevant to assessing world-similarity in the evaluation of counterfactuals. The apparent truth of MC is also why many now agree that we must appeal to causal independence in our evaluation of counterfactuals. The thought is that to get counterfactuals like MC to be ruled *true* on nearly any plausible analysis of counterfactuals, we must only count for similarity (or, alternatively, only "hold fixed") facts causally independent of the antecedent.

But appealing to causal independence in this manner is no trivial modification to our semantics for counterfactuals. For one, it means understanding counterfactuals in terms of causation, which rules out giving a reductive analysis of causation in terms of counterfactuals, and is, as such, a serious blow to the Humean conception of causation.

In the first paper I argue that MC is *false*, as is the thesis that the correct semantics of counterfactuals must appeal to causal independence (call the latter thesis the "causal independence thesis"). More generally, I *reject* the following principle:

Ahmed's Principle: For two actual events C and E, if C makes no causal difference to E then E would have still occurred even if C had not.[1]

First, I argue that every rationale one might reasonably have for thinking that MC (and Ahmed's principle) is true is based, at bottom, on implicit deterministic assumptions. I then show that when the causal independence condition is understood in the only way that is appropriate in this context, it ends up entailing that many unequivocally false counterfactual conditionals are true (and so the causal independence thesis should be rejected). In addition, I argue that other counterfactuals that have been taken to provide further support for the causal independence thesis fail to do so in each case.

If I am right that MC is false, this undercuts one of the most threatening objections to the counterfactual analysis of causation (i.e., the objection that the correct treatment of causation must itself invoke causal notions). But there are additional well-known and serious difficulties for the counterfactual analysis. In the second article I argue that there are important, widely-held assumptions about difference-making and its relationship to causation which are false, and which may be at least part of the source of some of these difficulties. I intend for this article to be the start of a longer project aimed at developing a reductive analysis of causation based on the alternative account of difference-making I begin to develop here.

In the third article I discuss the problem of reverse Sobel sequences – an alleged problem for the classic Lewis-Stalnaker semantics for counterfactuals. Sobel sequences are sequences of two counterfactuals in which the antecedent of the second counterfactual is stronger than the antecedent of the first and the consequent of the second is the negation of the consequent of the

[1] Arif Ahmed defends this principle in his (2011).

first. For instance "If Sophie had gone to the parade she would have seen Pedro Martinez (who was featured on a float), but if Sophie had gone to the parade and been stuck behind someone tall she wouldn't have seen Pedro Martinez on the float." It is sequences like these that motivated Lewis to reject a strict conditional semantics for counterfactuals in favor of his now orthodox variably strict conditional semantics. It is only the latter that can rule both counterfactuals in the sequence true in a fixed context. The problem of reverse Sobel sequences is that the two counterfactuals no longer seem consistent if the order of utterance is reversed: it seems there is something wrong with saying, "If Sophie had gone to the parade and been stuck behind someone tall she would *not* have seen Pedro, but if Sophie had gone to the parade she would have seen Pedro". The difficulty for the Lewis-Stalnaker semantics is that it cannot account for this difference between Sobel sequences and reverse Sobel sequences. In response to the problem, some theorists have argued that we should reject the classic semantics entirely. Others have argued that the infelicity of reverse Sobel sequences has a pragmatic explanation and should not be attributed to one of the counterfactuals being *false*. I argue that none of the extant approaches to the problem are right, and defend a novel solution to the puzzle. If I am correct, reverse Sobel sequences do not threaten the classic analysis. They do, however, give additional evidence for the thesis, forcefully defended by Alan Hájek, that most non-probabilistic 'would'-counterfactuals are false. This motivates placing a stronger emphasis on trying to understand probabilistic counterfactuals first and foremost.

MORGENBESSER'S COIN

Section I

Before flipping a fair, indeterministic coin, Susan offers Lucky the chance to bet on heads. Lucky declines. The coin lands heads. Consider the following counterfactual:

(M) If Lucky had bet heads he would have won.

Intuitively (M) seems true (assume that Lucky's bet would not have had a causal impact on the toss). I will refer to this intuition as the "ordinary intuition". This scenario, named Morgenbesser's Coin after Sidney Morgenbesser (who cites Michael Slote 1978), has had, and continues to have, an extraordinary influence on the literature on counterfactuals. I will cite some examples of its impact. The Lewisian (1973, 1979) truth conditions for counterfactuals are given below.

> A counterfactual "If it were that A, then it would be that C" is (non-vacuously) true if and only if some (accessible) world where both A and C are true is more similar to our actual world, overall, than is any world where A is true but C false. (Lewis, 1979, p. 465)

The similarity ordering is then given by the following similarity weighting:

> 1. It is of the first importance to avoid big miracles.
> 2. It is of the second importance to maximize the region of perfect match.
> 3. It is of the third importance to avoid small miracles.
> 4. It is of little or no importance to maximize the region of imperfect match. (Lewis, 1979, p. 472)

This ordering leaves open whether imperfect match of particular fact should count for nothing, or whether it should have relatively little weight. This reflects Lewis's own uncertainty:

> It is a good question whether approximate similarities of particular fact should have little weight or none. Different cases come out differently, and I would like to know why. Tichy and Jackson give cases which appear to come out right under [the analysis shown above] only if approximate similarities count for nothing; but Morgenbesser has given a case, reported in Slote ([1978]), which appears to go the other way. (Lewis, 1979, p. 465)

Despite the fact that it was Morgenbesser's coin that apparently motivated Lewis's inclusion of the fourth criterion (the first three criteria, by themselves, rule the counterfactual *false*) the similarity metric given above is nevertheless still unable to rule (M) *true*, even with the inclusion of (4). That is because although a (Lucky-bets-heads-and-coin-lands-heads)-world, call it w1, preserves the outcome of the coin toss—and thus, one aspect of imperfect match—a (Lucky-bets-heads-and-coin-lands-tails)-world, w2, preserves a different aspect of imperfect match: match in the outcome of the bet (i.e., in w2, like in the actual world, Lucky *loses*). Since, as Jonathan Schaffer points out, "either [match in coin toss outcome or match in bet outcome] might have the wider ramifications – for instance, either might inspire Nixon to press the button [resulting in nuclear war]" (2004: 303), we can easily revise our background story so that (1)-(4) rules (M) as clearly *false* (if w1 and w2 turn out to be equidistant from the actual world, (M) is also false).

In response to this apparent problem many proponents of the traditional possible-worlds framework have attempted to modify Lewis's similarity metric so that it can rule (M) as true, regardless of the consequences or ramifications of the toss's outcome. The consensus among those who endorse a modification to the similarity metric in response to (M) has been that it should be modified in the following way: only facts that are causally independent of the counterfactual's antecedent should be counted as relevant for comparative world similarity.[2] Let us call this the *causal independence thesis.* If only facts that are causally independent of the antecedent count toward the world similarity ordering (that is, if the causal independence thesis is true), then (M) is true.[3]

[2] See, e.g., Bennett (2003), Schaffer (2004) and Edgington (2004).

[3] Noordhof (2004) maintains that we can get (M) to be ruled true by only counting toward similarity facts *probabilistically* independent of the antecedent. The problem, as Schaffer (2004) has pointed out, is that although

That is because although the outcome of the toss is both causally and probabilistically independent of Lucky's bet, the outcome of his bet—i.e., whether he wins or loses—is not. On this proposal, Lucky-bets-heads-coin-lands-heads worlds are closer than Lucky-bet-heads-coin-lands-tails worlds, because the preservation of the toss outcome counts towards similarity, whereas the preservation of betting outcome does not.

Incorporating the causal independence thesis is no trivial revision to the similarity account. It means appealing to causation in our analysis of counterfactuals, which rules out giving a reductive analysis of causation in terms of counterfactuals, and is, as such, a serious blow to the broadly Humean conception of causation.[4]

The assumed truth of (M) has had a significant impact on the literature on counterfactuals in other ways as well. Some theorists have cited an ability to rule (M) *true* as evidence for alternative semantic accounts that depart from the Lewis-Stalnaker picture entirely. Hiddleston (2005), for example, has appealed to the apparent truth of (M) as evidence for his causal-model based theory of counterfactuals. Indeed, if it turns out that (M) is actually false, then Morgenbesser's coin will provide us with a counterexample to causal-model accounts of this sort.

Morgenbesser-style cases have also been cited as the reason to reject Dorothy Edgington's (1995, 2003, 2008) influential suppositional analysis of counterfactuals. On Edgington's view,

whether the coin lands heads or tails is probabilistically independent of Susan's toss, it seems that the outcome of the bet (that is, whether Lucky wins or loses) is *also* probabilistically independent of the toss - and yet we don't want to count that as relevant for similarity. Noordhof (2005) has a way around this, although it requires accepting other of his controversial commitments. Nonetheless, each of the arguments I give here works just as well against the view that only facts probabilistically independent of the antecedent should count toward similarity.

[4] Humeans take causation to reduce to something like mere counterfactual dependence.

> …confidence in [a] counterfactual expresses the judgment that it was probable that B given A, at a time when A had non-zero probability, even if it no longer does; and even if you do not now have a high degree of belief in B given A. (Edgington 1995: 265)

According to Edgington, belief in the counterfactual A > C is *acceptable* only if belief in the indicative conditional A→C was acceptable at some prior time when A had a non-zero probability.[5] For example, it is acceptable to believe the counterfactual that it is very likely that Fred would have been cured had he undergone the operation just in case, at a time prior to Fred's decision not to have the operation, it would have been acceptable to believe that it is very likely that Fred will be cured if he undergoes the operation (Edgington 2004: 9).

Bennett (2003) rejects this account on the grounds that it seems to rule belief in Morgenbesser counterfactuals *unacceptable.*[6] That's because, prior to the Lucky's bet it would *not* have been acceptable to believe that A→C (if Lucky bets heads he will win).

Given how much hangs on the truthvalue--or the acceptability status[7]--of (M), it has been a mistake, I think, to simply take the ordinary intuition for granted. Especially since, as I will now argue, there are good reasons to question the assumption that Morgenbesser counterfactuals are true.

Section II

There are two different ways to think about Morgenbesser's coin. Thought about one way, it seems evident that the counterfactual is true. Reasoning about the counterfactual in a second

[5] For Edgington, conditionals with false antecedents are not propositions and do not have truthvalues. Conditionals with false antecedents are deemed as either *acceptable* or *unacceptable*, depending on their fittingness for belief.

[6] Edgington (2004) amends her account in response to the objection so that it can rule Morgenbesser counterfactuals acceptable. Her amendment is problematic, however, as Phillips (2007) shows.

[7] Henceforth for simplicity I drop the talk of *acceptability* and speak about counterfactuals under the assumption that they are propositions with truthvalues. Whether they are propositions or not has no bearing on what is said here.

way, however, leads to the opposite conclusion. I suspect that those who immediately judge (M) to be true do so because they are thinking about the counterfactual in the first way. The problem is that, as we shall see, the first way is unmotivated in an indeterministic context. In the next section of the paper I will give an additional argument for rejecting the causal independence thesis. If the causal independence thesis is false, there is no plausible way for (M) to be ruled true on any of the prominent analyses of counterfactuals. Before getting to this argument, however, we should first take a look at the two different ways one might reason about (M).

The first way to reason about (M) is as follows. Given that the coin actually lands heads, and that, had Lucky bet heads, his bet would have had no impact on the toss, his bet would have made no difference to the *outcome* of the toss, either. Therefore, the coin would have (*still*) landed heads. Arif Ahmed explicitly reasons in this way in defense of the ordinary intuition about (M) (2011:80):[8]

> (1) If C makes no difference to an actual event E then E would still have occurred even if C had not (premiss).
> (2) C makes no difference to any actual events to which it is causally irrelevant (premiss).
> (3) [Lucky's not betting heads] is causally irrelevant to the [outcome of the toss] (premiss).
> (4) Therefore [M] is true.

Here, on the other hand, is the second way to reason about (M). Consider how Alexander Pruss (2003) describes our "ordinary thinking" about counterfactuals:

> In the case of our ordinary thinking about counterfactuals, it is natural to locate, with some vagueness, the first event in respect of which the counterfactual world is supposed to diverge from the actual world, and then to consider how the divergence causally propagates as a result of this event.

[8] Ahmed's argument is, as far as I know, the only argument made in defense of the ordinary intuition about (M). In all other cases, that (M) is true is simply taken for granted.

Counterfactual semantic models generally aim to capture something approximating how we ordinarily understand counterfactual assertions. And we ordinarily understand a counterfactual assertion as asserting something like the following: if the antecedent (which in most cases is actually false) were true, and if the world prior to the antecedent were otherwise approximately the same (with only minor differences required to make the antecedent true), then the consequent would follow. We make the required changes *prior* to the time of the antecedent's obtainment, and let things unfold, as they will, from there. If the consequent obtains, the counterfactual is true. If the consequent does not obtain, the counterfactual is false. Let's start with this rough and ready picture and see what happens when we use it to evaluate (M). First, we "go back" to the time, prior to the antecedent, when the (minimal-possible) changes need to be made for the antecedent to obtain: in this case, for Lucky to decide to bet heads. Call the moment of the first required change the *fork*. We then let the subsequent events unfold.

In a *deterministic* world, since Lucky's bet plays no causal role in the outcome of the coin toss, the coin would still land heads (we can assume that none of the minor changes required to make Lucky decide to bet heads would themselves have a causal impact on the coin toss, either). Since the coin toss is stipulated to be indeterministic, however, if we let the sequence of events play out from the point at which Lucky bets heads, there is no guarantee that the coin will still land heads. It could land either heads or tails, despite not being influenced by Lucky's bet. Unlike in the deterministic case, in the indeterministic case we'd have to artificially *hold* the outcome of the toss fixed for (M) to be true.

The second way to reason about (M), then, is this. The change in the fork that results in

Lucky betting heads takes us on a different post-fork path – or, in world talk, it takes us to a different world. Had Lucky bet heads the world would not have been the actual world ('actual world' should be understood as a rigid designator, here). And, since the toss is indeterministic, in a different world the coin could land either heads or tails: there's no reason to think that the outcome in the counterfactual world at which Lucky bets heads would necessarily match the outcome in the actual world. Since the coin could have landed heads or tails at the relevant world or worlds where Lucky bets heads, (M) is false.

This way of reasoning gives us a way out of Ahmed's argument. Ahmed's first premise says that if C makes no difference to an actual event E then E would still have occurred even if C had not. Substituting in the actual fact that Lucky did *not* bet heads for C, and the actual fact that the coin landed heads for E, premise (1) says that had C not occurred—that is, had Lucky *bet* heads—the coin would have (still) landed heads. On our alternative picture, this premise is false. E can fail to occur following C not occurring despite the fact that C does not itself make a difference to E. That is because the change in the fork, which results in C not occurring, takes us to a different world. And at a different world, the coin could land heads or tails. C need not be causally connected to E for a change in C to correspond with a possible change in E.

So which way of thinking about (M) is right? Clearly Ahmed has ordinary intuition on his side. Each premise of his argument, including premise (1), is intuitively appealing, at least at first sight. And of course, most have the intuition that (M) is true. It is clear that Ahmed's first premise *is* true in a deterministic context. If C makes no causal difference to E and E is the result of a deterministic sequence, then E would have still occurred even if C had not. Since the other premises in the argument seem acceptable, it would be entirely unsurprising for his argument to *seem* sound, even if it were not. After all, we are not used to thinking indeterministically: it is perfectly reasonable to expect that our intuitions have an underlying deterministic influence.[9] We shall now investigate whether there is any legitimate reason to think that Ahmed's way of reasoning is right in the indeterministic case.

Let us begin by identifying how someone defending the truth of (M) would object to my alternative proposed way to reason about (M). (I've already said where Ahmed's reasoning goes wrong, on my view.) Someone endorsing the truth of (M) would presumably say the following. It is true that at each distinct world at which the indeterministic coin (or more precisely, one of its counterparts) is tossed, it could land either heads or tails. Nonetheless, since the coin lands heads in the actual world, other worlds at which the coin lands heads are *more similar* to the actual world than are worlds at which the coin lands tails. If so, it does not matter to the evaluation of the counterfactual that half of the coin-toss-worlds are coin-lands-tails-worlds. The coin-lands-heads-worlds are not among the most similar to the actual world.

We've already seen that although the worlds at which Lucky bets heads and the coin lands heads are more similar to the actual world in one respect (they match in toss outcome), the worlds where Lucky bets heads and the coin lands tails are more similar in a different respect: namely, in these worlds, like in the actual world, Lucky loses the bet. It is at this point that the defender of the truth of (M) must appeal to causal independence: she must hold that only facts causally independent of the antecedent count for similarity. But why should we think this right? Notice

[9] Phillips (2007) makes a similar point.

that Ahmed's argument cannot be of help to her, here. Ahmed's argument cannot help to explain why only facts causally independent of the antecedent should matter for similarity, since the first premise of his argument *already presupposes* that the causal independence thesis (or something else that can take its place) is right. Ahmed's first premise is only true if (only) facts causally independent of the antecedent count for similarity.[10] And it is precisely this that is at issue in the decision over which of the two ways is the right way to reason about (M). To tip the balance in favor of Ahmed's way of reasoning we need some principled reason to think that *even in the indeterministic case*, there is something special about facts that are causally independent of the antecedent: something that makes it such that these facts, and no other ones, should count toward similarity.

On the contrary, there is reason to think that in the indeterministic case there is nothing special or significant about facts that are causally independent of the antecedent. We can see this if we consider why these facts might be thought to be special (and why they *are* special under determinism). In his (2004), Schaffer concisely states why causal dependence and independence seems to matter for similarity. He writes the following:

> Here is one way to express [the idea of invoking causal independence]: only match among those facts *causally independent of the antecedent* should count towards similarity. After all, if outcome o causally depends on p or ~p, then o should be expected to *vary* with p or ~p – its varying should hardly count for *dis*similarity." (2004: 305, his emphasis)

In general, if some effect, E, causally depends on some cause, C, then it is expected that E will vary with C. And if E would vary with C in the actual world, we should expect E to also vary

[10] That is, unless some better way to distinguish similarity in toss-outcome and similarity in bet-outcome can be found. This seems unlikely.

with C in the nearest possible worlds: if anything, that E varies with C in some world, w1, makes w1 *more* similar to the actual world than is a different world, w2, where E does not vary with C (since E's failure to vary with C suggests that there isn't the same relation of causal dependence at that world).

For this reason it makes sense to think that we should hold fixed only what is causally independent of the antecedent, and allow that which is causally dependent on the antecedent to vary with the antecedent as it will. But notice that this motivation for the causal independence condition does not extend equally well to all cases. For example, it does not extend to worlds where the probability of the effect is *exactly the same,* given the cause. For in that case, there is no reason to think that varying the cause should necessarily result in a variation in the effect.

Suppose that Susan tosses a fair, indeterministic coin in w1, and Lucky tosses a fair, indeterministic coin in another world, w2, that is otherwise just the same. Suppose that at both worlds the coin comes up heads. Since whether it is Susan or Lucky who tosses the coin does not matter at all to the probability of the outcome, there is no longer the same motivation for thinking that the outcome being the same in both worlds *shouldn't* make the worlds more similar to one-another than to a third world where the outcome is different. Indeed, the fact that the outcome is the same in both worlds should now, arguably, count *in favor* of their similarity if imperfect match is relevant for similarity at all.[11] (This is a good reason to deny that imperfect match is relevant for similarity at all!)

As such, it remains entirely mysterious why we should think that *even in an indeterministic*

[11] I am indebted to Carolina Sartorio for this way of putting the point.

context there is something significant about facts that are causally independent of the antecedent which can justify holding only these facts fixed. On the other hand, there is a perfectly good explanation for why many of us have the *intuition* that only such facts count for similarity: in a deterministic setting this is the case, and we are not frequently reasoning about genuinely indeterministic processes. Indeed, it is reasonable to think that our experience at the macro level is as if the world were deterministic. In the next section I show that when the causal independence thesis is understood as it should be, it ends up entailing that many unequivocally false counterfactuals are true. This gives us additional, positive reason to reject it.

Section III

I will now argue that there is good reason to reject the causal independence thesis. If I am right that it is false, we will be left with no plausible way to account for how (M) could be true on a similarity framework.[12]

In this section my focus is not directly on the truthvalue of (M). Here I investigate whether the causal independence thesis can give us the correct verdict for other counterfactuals. If it turns out that the thesis is wrong, then the best (and indeed, so far, *only*) proposals for ruling (M) as true

[12] Nor on a premise semantic framework. According to the premise semantic model (which can be traced to Goodman (1946) and has been defended by Barker (1999) and Hiddleston (2005), among others) a counterfactual is true just when the consequent follows from the antecedent and some specified subset of actual facts. Like possible world accounts, premise semantic accounts have a difficult time predicting that Morgenbesser counterfactuals are true. That's because if we hold fixed only the *prior-to-his-betting* facts compatible with his betting heads, as seems the natural thing to do, the coin toss could come up either heads or tails. For (M) to be true we need to hold fixed the post-antecedent fact that the coin landed heads. Premise semanticists have proposed a solution parallel to Schaffer's own: hold fixed only facts causally independent of the antecedent, regardless of their temporal relation to the antecedent. Since the coin landing heads is causally independent of Lucky betting heads we hold the result of the coin toss fixed (but not the actual fact - causally *dependent* on the antecedent - that Lucky did not win the bet), and (M) comes out true. This variation of the causal independence thesis faces all the same difficulties.

will have been refuted.[13] This would not bode well for the ordinary intuition about (M).

Unfortunately for the ordinary intuition there is, I will now argue, a counterexample to the causal independence thesis. Suppose that Lucky bets heads and Susan tosses a coin and it lands tails. We shall stipulate that (i) the result of the coin toss is indeterministic and (ii) each time the coin is tossed it always has the same chance of landing tails (for simplicity we can assume that, that chance is 0.5). Now consider the following counterfactual (discussed, for very different purposes, by Bennett (2003)):

(B) If Lucky had been the one to toss the coin, it would have landed tails.

I think it will be agreed that (B) is false. If Lucky had been the one to toss the coin there's no saying whether it would have come up heads or tails. (Note that for our purposes we need not assume that everything about the coin toss other than who actually tossed it is held fixed. We are free to assume that had Lucky been the one to toss the coin he would have tossed it differently than Susan did. On the other hand, the case also works if we *do* assume that Lucky would have tossed the coin just as Susan did. What matters, for my purposes, is that it has a 0.5 chance of landing tails regardless of the particular details of how it is tossed.) Does the causal independence thesis correctly predict that (B) is false? At first sight, it might seem to. Indeed, counterfactuals like (B) have been used as further evidence *for* the causal independence thesis, precisely because it

[13] For simplicity I focus exclusively on the causal independence thesis, although once again my argument can be used against the probabilistic version of the thesis just as well. The probabilistic version says that all facts probabilistically independent of the antecedent ought to be held fixed when evaluating a forward-tracking counterfactual.

seems to rule Morgenbesser counterfactuals true while still being able to rule (B) false.[14] To arrive at the conclusion that (B) is false assuming the truth of the causal independence thesis, one reasons as follows. Since the coin landing tails is *not* causally independent of Lucky tossing the coin, we do not hold the outcome of the coin toss fixed (alternatively, on a possible worlds framework, the outcome of the coin toss does not factor into the world-similarity ordering). Since the outcome is not held fixed, and since the toss is indeterminate, the coin could land either heads or tails, and (B) comes out false (as it should).[15]

It appears, therefore, that invoking causal independence in the way Schaffer and others have done to account for Morgenbesser's coin also gives us the correct truthvalue for (B). I want to suggest that the reasoning leading to this conclusion is erroneous, however, and that in fact, the causal thesis commits us to the *truth* of (B). Since (B) is indisputably false, if I am right that the causal thesis rules (B) true we ought to reject the causal thesis.

But why would the causal thesis rule (B) true? The causal thesis rules (B) true if the outcome of the coin toss is in fact *not* causally dependent upon Lucky tossing the coin. And indeed, I claim, it is not. The outcome of the coin toss is causally *independent* of Lucky tossing the coin. The reason is simple. Since the coin toss is indeterministic, the outcome of the toss is causally independent not only of *who* tosses the coin, but also of *when* the coin gets tossed, *where* the coin gets tossed, *how* the coin is tossed. Since the outcome is causally indeterminate, it is

[14] Won (2009), for instances, uses (B) to support the causal independence thesis in just this way, although he does not explicitly stipulate that the coin has the same probability of landing heads whether it is Lucky or Susan who tosses it. If he is not making this assumption implicitly then his case is importantly different from mine.

[15] According to the Lewisian analysis of counterfactuals (B) also comes out false if the outcome of the toss does not count toward similarity, since the Lucky-toss-worlds in which the coin lands tails will be no closer than the Lucky-toss-worlds in which the coin lands heads.

causally independent of the entire conjunction of facts that together constitute every describable aspect of the toss.

But if this is right then why was it so easy to be convinced that the outcome of the toss is causally affected by Lucky being the one to toss the coin rather than Susan? It's not at all surprising that it was, considering that there is *one sense* in which the outcome of the toss is causally affected by who tosses the coin. It's just not the sense that matters. For although *whether* the coin lands heads or tails is causally independent of Lucky having been the one to toss the coin, *that* the coin lands *either* heads or tails is not. If it would have landed heads, then in virtue of Lucky being causally responsible for it landing anything at all (that is, either heads or tails), he would have been causally responsible for its landing heads: but only because it *did* land heads instead of tails (for *a*-causal reasons). And if it would have landed tails, then in virtue of being causally responsible for its landing anything at all, Lucky would have been causally responsible for its landing tails: but only because it *did* land tails instead of heads (for *a*-causal reasons).

In other words, had Lucky tossed the coin—still keeping in mind, of course, that we are assuming that the outcome of the toss is genuinely indeterministic—Lucky's toss would *not* have played a causal role in determining *which* side of the coin landed face up. However his toss *would* have played a causal role in bringing about either that it landed heads, or, alternatively, that it landed tails, *depending on which happened*.

One way to frame this distinction is to distinguish between two possible uses of the term '*outcome*'. Had Lucky been the one to toss the coin, Lucky would have been causally responsible for the outcome if 'outcome' is understood in what I will call the *non-comparative sense* ("outcomeNCS"). Whatever the outcome would have been, Lucky would have been causally responsible for it in virtue of being causally responsible for the coin landing either heads-up, or tails-up, at all. However, Lucky would *not* have been causally responsible for the outcome in the *comparative sense* ("outcomeCS"). He would not have been causally responsible for the coin landing heads *rather than tails*, nor for it landing tails *rather than heads*.[16] He would not have been causally responsible for which of the two outcomes would have obtained, had he been the one to toss the coin. After all, since the coin toss is indeterministic, which side lands face up is *random*.

If this is right then although the outcomeNCS is causally dependent on Lucky tossing the coin, the outcomeCS is not. And just a bit of reflection about (B) reveals that it is outcomeCS that is at issue. When assessing (B), whether the coin would have landed *one of* either heads or tails had Lucky tossed the coin is certainly not what is at issue. What is at issue is *which* side would have landed face up, had Lucky been the one to toss the coin. Would it have been tails rather than heads (as the counterfactual claims) or heads rather than tails?

To bring this out more clearly it may be helpful to compare our scenario to another

[16] I use the term "comparative" rather than "contrastive" to try to avoid giving the impression that I am presupposing a contrastive view of causation like the one given in Schaffer (2005). Certainly my argument is compatible with Schaffer's account, which holds that causation is not a binary relation between cause and effect but is instead a "*quaternary*, *contrastive* relation: c rather than C* causes e rather than E*, where C* and E* are nonempty sets of contrast events." (297, his emphasis) Put in the terms of his analysis, it might be appropriate to understand "outcomeNCS" as picking out the contrast between the coin landing *something*—i.e. either heads or tails—rather than nothing and "outcomeCS" as picking out the contrast between the coin landing heads rather than tails (or alternatively tails rather than heads). This is a convenient way to capture my distinction, so there is potentially some benefit for my argument if his analysis, or another like it, is the right one. Nevertheless I don't want to take a stand on this issue either way so long as the causal theorist who rejects contrastivism is able to somehow capture the distinction that I have called "outcomeNCS" and "outcomeCS". If a theory of causation is unable to say that the person who tosses the indeterministic coin is causally responsible for the disjunction <the coin lands heads or tails> (and thus, for it landing heads if that is what happens) but is not causally responsible for whether it lands heads rather than tails or vice versa – or, to take an even clearer case (about to be discussed), if a theory of causation is unable to say that the terrorist who sets an indeterministic time bomb is causally responsible for the bomb detonating (assuming it does) but is not causally responsible for whether or not the bomb detonates after it is set - then this, it seems to me, would be a serious problem for that theory.

scenario involving probabilistic causation. Imagine that a terrorist is considering setting an indeterministic time bomb which, if set, has a 50% chance of detonating. In this scenario, as I've described it, there are at least the following three contextually salient possibilities:

(i) The terrorist sets the bomb and it detonates.
(ii) The terrorist sets the bomb and it does not detonate.
(iii) The terrorist does not set the bomb (and it does not detonate).

Suppose the terrorist decides to set the bomb and it detonates. By setting the bomb, the terrorist eliminates possibility (iii). For the purposes of this example let us adopt a probabilistic understanding of indeterministic causation according to which *to cause* is understood as something like *to make more likely to happen.*[17] The terrorist is causally responsible for the detonation of the bomb in the following sense: by setting the bomb he eliminates possibility (iii), and thereby raises the probability of (i) to 0.5. This is all that is required for the terrorist to be causally responsible for the detonation in the *non-comparative* sense. (It is the *non-comparative* sense of 'outcome', I take it, which is what is relevant in most moral contexts involving probabilistic causation.) Had (i) obtained, the terrorist would be causally responsible for (i) in virtue of having eliminated possibility (iii) (or alternatively, if it is preferred, we can instead say that the terrorist would be causally responsible for (i) in virtue of his bringing about the possibility that (i) could obtain). But here is the crucial point: the terrorist is *not* causally responsible for which of the remaining two possibilities—i.e., either (i) or (ii)—obtains if he sets the bomb. Given that he sets the bomb, whether it goes off or not is entirely up to chance.

[17] I choose this conception of indeterministic causation for simplicity and because it is, as far as I can tell, the most widely held account. Not much should be made of this choice, however. My argument could be made just as well in terminology consistent with most alternative conceptions of indeterministic causation.

Let us return now to (B). In most contexts in which (B) might be uttered, in the counterfactual scenario in which Lucky tosses the coin there are *two* salient alternatives.[18] Either:

(iv) The coin lands heads. Or,
(v) The coin lands tails.

Unlike in the terrorist scenario in which there was the possibility that the terrorist might not set the bomb at all, in the counterfactual scenario there is *not* the possibility that Lucky might not toss the coin. The antecedent tells us that in the counterfactual scenario Lucky tosses the coin. Since it is a given, in the counterfactual scenario, that Lucky tosses the coin, the truthvalue of (B) depends only on *which* of the two alternatives, (iv) and (v), are true, given Lucky's toss. And just as the terrorist would not have been causally responsible for which of (i) or (ii) would have obtained had he set the bomb, Lucky would not have been causally responsible for which of (iv) or (v) obtained, had he tossed the coin. Since, as I have argued, *whether* the coin lands heads or tails is causally independent of Lucky being the one to toss the coin, the causal thesis tells us to hold the actual fact that the coin landed heads rather than tails, fixed. If we hold fixed that the coin lands heads, however, then (B) comes out true. Since it is evident that (B) is in fact *false,* it can be concluded that the causal thesis gives the wrong ruling for (B).

Section IV

I now attempt to preempt some possible objections. It might be tempting to think that an important difference between (B) and (M) is that were Lucky to toss the coin rather than Susan, he

[18] Of course, in some contexts there might be more than two salient possibilities. For instance, it may be appropriate to consider the possibility that the coin could have landed neither heads or tails had Lucky tossed it: by freak chance, perhaps it could have landed standing up on its edge. This does not make a difference to which kind of *outcome* is the relevant one, however. What matters is that (B) does not leave open the possibility that Lucky did not toss the coin at all in the counterfactual scenario.

would be initiating a causal sequence that is distinct from the actual causal sequence – i.e., the one initiated by Susan. In contrast, in the case of Morgenbesser's Coin, the relevant causal sequence intuitively seems to be the *same* causal sequence, whether Lucky does or does not place the bet. The objection, then, might be put like this: a distinct "coin-tosser" entails a distinct coin toss. In contrast, Lucky placing a bet that is causally independent of the toss does not entail a distinct toss. And of course, for each distinct coin toss trial we should expect the probability that the coin lands heads to start anew.

In other words it might be thought that the relevant difference between the two cases is that the causal chain terminating in Susan's coin toss and outcome is distinct from the causal chain terminating in Lucky's toss and outcome. In contrast, in the Morgenbesser scenario, although the world in which Lucky bets heads is distinct from the world in which he does not bet heads, the tosses share the same (i.e. counterpart[19]) causal chains.

The problem with this way of thinking, however, is that it conflates outcomeCS and outcomeNCS. While it makes sense to speak of coin tosses and their NCS-outcomes as being the result of causal chains, it does not make sense to speak this way about the outcomeCS. No causal chain determines which side of the coin lands face up. Recall the terrorist scenario again. If the terrorist at issue sets the bomb, he initiates a causal sequence that results in the bomb being in a state such that it might detonate. If it detonates he is causally responsible for this. Nevertheless, nothing is causally responsible for whether the set bomb detonates or not. No causal chain selects between (i) and (ii).

[19] Henceforth I omit this clarification.

The point can be made in a different way. There is a difference between the scenario where Lucky bets heads (and Susan tosses the coin) and the scenario where it is Lucky who tosses the coin rather than Susan. That difference is that Lucky initiates a distinct causal chain (which terminates in the coin landing face-up on either one side or the other) in the second case but not in the first. Is this difference a relevant one? It is in *deterministic* contexts. (You might recognize where this is going.) If the causal chain resulting in the coin landing heads is deterministic, then at each distinct world with the same causal chain, the coin lands heads. In contrast, at worlds at which a different person tosses the coin initiating a distinct causal sequence, the coin could land either heads or tails. So an explanation for why *intuitively* it seems to be relevant whether or not there is a distinct causal process leading to the outcome is readily available: if the toss were deterministic, this distinction *would* be relevant. If the outcome of the toss is indeterministic, however, then even in worlds with chains identical to the actual one, the coin has only a 0.5 chance of landing heads. As such, there's no longer the same justification for thinking that having the same causal chain makes a difference to whether match in outcome matters for similarity or not. It seems that either it should matter in both cases (that is, in the Morgenbesser scenario and the (B) scenario), or else it should matter in neither: for in *neither* case does the causal chain fix the outcome, or even make one outcome more likely than the other. And if the probability of heads is the same at the possible worlds at which Lucky tosses the coin as it is at the worlds at which Susan does, then it is (once again) entirely unclear why we should think that there is anything significant about identity in causal chains which makes it such that, even in indeterministic contexts, match in outcome is relevant for similarity just in case the causal chains leading to the outcomes are the same.

Edgington (2014) has an independent argument in support of holding fixed facts concerning times later than the antecedent-time in the evaluation of counterfactuals. While this argument does not directly support the causal independence thesis, if it is successful it takes us a lot of the way there: if post-antecedent facts should be held fixed at all, it is much more plausible that only facts causally independent of the antecedent should be (at least given some appropriate understanding of causal independence). Edgington argues that our use of counterfactuals to make inferences in the empirical sciences requires that we hold fixed facts concerning times later than the antecedent-time. Consider her example, below.

> A long time ago, a volcano erupted. It was a slow eruption, the lava creeping onwards slowly. At that time, it was very likely that the lava would eventually submerge valley A, but valley B would not be affected--given the lie of the land. However, in the unlikely event of an earthquake of a particular kind at an appropriate time, the path of the lava would very probably be switched away from valley A, towards valley B. As a matter of fact, this is what happens. Along comes our geologist, centuries later, making his inference about the eruption. He has already found out about the earthquake. "That volcano must have erupted", he concludes, "For there is lava in valley B and not in valley A; and, given what I know about the earthquake, that is just what one would expect to find if that volcano had erupted".

Suppose there was a second volcano whose potential eruption, at the time in question, presented much more danger to valley B; but in the unlikely event of the earthquake, its lava would probably be diverted elsewhere. Only with hindsight (knowing of the earthquake) is one justified in thinking that if the second volcano had erupted, valley B would not have been submerged; and if the first had erupted, it would have been submerged. (2004:24)

Because the geologist accepts the counterfactuals

(i) If the *first* volcano had erupted, valley B would have (probably) been submerged, and
(ii) If the *second* volcano had erupted, valley B would have (probably) *not* been submerged,

the geologist can (correctly) infer from the fact that valley B was submerged that it was the first volcano which (probably) erupted. But neither (i) nor (ii) is acceptable if the fact that the earthquake occurred is not held fixed. And the earthquake took place *after* the volcano erupted – that is, after the time of the counterfactuals' antecedents.

Examples like the one above seem to show that post-antecedent facts must, at least in some circumstances, be relevant for similarity if counterfactuals are to successfully serve their important function in inferential reasoning. But there are a couple of things to note about the example. The first is that, in the context in question, (i) and (ii) can plausibly be understood as elliptical for (i') and (ii'):

> (i') If the first volcano had erupted *and the earthquake occurred*, valley B would have (probably) been submerged, and
>
> (ii') If the second volcano had erupted *and the earthquake occurred*, valley B would have (probably) *not* been submerged,

The earthquake is known to have occurred – what is unknown, to begin, is which volcano erupted. When uttering the counterfactuals the geologist might not refer to the earthquake explicitly, but what he is trying to communicate seems to be accurately captured with (i') and (ii'): it is just being *assumed* that the earthuake would have occurred had either the first volcano, or the second volcano erupted, since what is in question is which volcano erupted *given* that the earthquake occurred.[20] In fact, if (i) and (ii) are *not* elliptical for (i') and (ii') respectively, then Edgington's example can be

[20] Note that this is importantly different from the context in which Morgenbesser's Coin is uttered, since there what is in question is not whether Lucky bet heads but what the outcome of the toss would have been if he had bet heads. If what was at issue, in the context, was whether Lucky bet heads, and if the actual fact that the coin landed heads would somehow give us some indication of the answer to that question (for instance, if we knew Lucky won and were reasoning in this way: if Lucky had bet heads and the coin had landed heads, he would have won. He did win and the coin landed heads, so he must have bet heads), then I think it likewise plausible that the relevant counterfactual would actually be <If Lucky had bet heads *and the coin had landed heads*, he would have won>.

used as a *counterexample* to the causal independence thesis. That is because when (i) and (ii) are used by the geologist in the context described above, the fact that the earthquake occurred should be held fixed *whether the earthquake is causally independent of both of the volcanoes or not*. Suppose that the eruption of the first volcano caused the earthquake to be more (or less) likely to occur. In this case the earthquake is not causally independent of the antecedent of (i) and yet, that the earthquake occurred must still be held fixed for (i) to come out true (or *acceptable,* in Edgington's terms). And (i) still needs to be true (or acceptable) for the geologist to use the counterfactual to make his inference – an inference which is just as good whether the earthquake is causally independent of the first volcano or not. And of course, the causal independence thesis says that *only* facts causally independent of the antecedent are to be held fixed in the evaluation of a counterfactual. So if to evaluate the counterfactual correctly the fact that the earthquake occurred needs to be held fixed even if it is not causally independent of the volcano, then the counterfactual functions as a counterexample to the causal independence thesis. Now if, as I suggest, (i) is elliptical for (i'), the counterfactual is not a counterexample to the causal independence thesis. However, in that case the counterfactual also does not show what Edgington claims that it shows, namely that "we need to take into account actual facts concerning times later than the antecedent-time" (2014: 24).

I have argued that contrary to initial appearances there is no principled way to distinguish how we ought to treat (M) from how we ought to treat (B), the latter of which is clearly false. There is *no* justification for thinking that in cases where (a) the outcome of the toss is indeterministic and (b) facts about how the coin is tossed (e.g., who tosses the coin) make no difference to the objective chance of each possible outcome – match in outcome only matters for similarity if the causal chains are the same. If the choice is between believing that match in the outcome of the toss matters for similarity in *both* cases or in neither case, the choice is clear: match in outcome is not relevant for similarity, even in a case like Morgenbesser's coin.[21]

[21] There are a number of reasons why the alternative choice is unacceptable. That (B) is clearly false is only one of them. Committing to match in outcome counting for similarity even when the two causal sequences are distinct will quickly lead to a proliferation of facts apparently mattering for similarity that shouldn't. If match in outcome matters for similarity even if the causal chains are distinct, then presumably match in outcome is relevant not only if different people toss the coin, but also if the coin is tossed at different times, or at different places. Maybe match in outcome even matters for similarity if the coins are different coins (or is there some justification for requiring that at least the coin be the same coin, even if what happens to it can vary?) Once we allow for outcome in coin toss to be relevant for similarity regardless of who tosses the coin, it's hard to see what principled way there could be to exclude a whole host of other unacceptable facts from counting.

CAUSATION AND DIFFERENCE-MAKING

Intuitively, causation seems to be intimately connected with difference-making. A cause, it seems, should make some sort of difference to its effect. It is has proven to be very difficult, however, to specify exactly how causation and difference-making are connected. It is widely held that it is sufficient for an event, C, to be a cause of some distinct event, E, if E would not have obtained had C not obtained: or in other words, if E's obtainment counterfactually depends on the obtainment of C. Call this the *difference-making principle* (DMP). DMP does not yet get us an analysis of causation, however, since it not necessary that E counterfactually depend on C for C to be a cause of E. In cases of redundant causation, C can be a cause of E despite the fact that C's obtainment makes no difference to whether or not E obtains. For example, suppose that Suzy and Billy each throw a rock at a window. Suzy's rock hits the window first and shatters it. We want to say that Suzy's throw was a cause of the window shattering even if the window would have still shattered (by Billy's rock) had Suzy not thrown hers.

There have been many attempts to strengthen DMP in order to come up with a necessary condition for C to be a cause of E. So far, none of these attempts have been entirely successful. Some have decided that it may be time to give up on trying to give a reductive analysis of causation either in terms of causes as difference-makers or otherwise. I think it is not yet time to give up. This is because I believe that there are important, widely-held assumptions about difference-making which are false. My hope is that if we can improve our understanding of what kind of difference-making is relevant for causation, we will be in a better position to determine just how difference-making is related to causation, and to evaluate the prospects for giving an account of the latter in terms of the former.

In this paper I am beginning my investigation into difference-making and its connection to causation. This is the start of what I hope will be a larger project. I will not be offering a positive account of difference-making or of its relationship to causation. My aim here is more modest: to make some progress toward identifying the best analysis of difference-making. I attempt to do this by taking a critical look at two compelling extant proposals on how best to understand difference-making: those of Boris Kment (2010) and Carolina Sartorio (2005). I will argue that neither of these accounts is ultimately successful. As we will see, identifying where each goes wrong will prove instructive.

I. I begin by examining Kment's recent account of difference-making and its relationship to causation. Kment focuses on two distinct phenomena. One is that there seems to be a very close connection between counterfactuals and causation. The second is that causation also seems to be closely tied with nomic determination. The *determination idea* is the idea that, assuming causal determinism, for any given effect, E, the set of all of its causes conjoined with the laws of nature must determine that E obtains. Thus, if all of the actual causes of E obtain, then E is nomically determined to obtain as well.

Kment points out that it is noteworthy that causation seems to be so closely tied to both determination and to difference-making. After all, the two concepts appear to be quite different. Difference-making, it seems, provides a sufficient condition for one event to be a cause of another. The determination idea, in contrast, captures what seems to be a necessary condition for a set of events to jointly cause E. Kment offers a way to explain why and how causation relates to both, and how determination and difference-making are related to each other. He argues that the best explanation for why causation intuitively seems to be so closely tied with difference-making, and thus with counterfactual dependence, is *not* that causation consists in some pattern of counterfactual dependencies. Rather, the role that difference-making and counterfactuals play in our causal thinking is explained through their ability to provide a useful test for our causal claims. And what justifies us in using counterfactuals to test causal claims is the determination idea. Let us take a look at how this works.

The "determination idea" can be captured by a principle that Kment calls "(D/d**)". (D/d**) gives a necessary condition for a set, S, to contain all the causes of some effect, E.

(D/d**) Those causes of E that obtain at *t* jointly nomically determine E. (2010: 95)

According to Kment, it is (D/d**) which justifies our use of counterfactuals to test causal claims. "Patterns of difference-making, like those we study in scientific experiments and counterfactual reasoning, are *not* what makes causal claims true" (my emphasis), he writes. "They merely provide a useful test for causal claims." Counterfactual reasoning is a useful heuristic for testing causal claims, and this is explained by (D/d**). What exactly is the connection between (D/d**) and counterfactual reasoning which explains why the latter can provide a good heuristic for testing causal claims? Kment explains that necessary conditions – like (D/d**) – can give rise to sufficient conditions with the help of the *method of elimination*. Here is how this works. Suppose N is a necessary condition for A. The method of elimination works when you know that there is some A and that, in addition, x and only x is N. In that case, you can infer that since nothing other than x is N, nothing other than x can be A. Since there is an A, x must be A.

For example, suppose that being in the garden at *t* is a necessary condition for being the murderer. The detective is aware of this necessary condition. She also knows that *someone* is the murderer. She doesn't know anything about the butler's whereabouts, but she knows the whereabouts of all other potential suspects, and knows that no one else was in the garden at *t*. Using the process of elimination the detective can infer that the butler must have been in the garden at *t* (since someone was). Since she knows that (i) no one but the butler meets the necessary condition to be the murderer, and (ii) there is a murderer, she can conclude that the butler is guilty. To put it slightly differently, (i) and (ii) conjoined provide a sufficient condition for the butler to be the murderer.

As we've just seen, with the help of the process of elimination the detective can use the necessary condition for being the murderer to derive a sufficient condition for being the murderer. Let's see how we can use this procedure to derive a sufficient condition from the necessary condition (D/d**). (D/d**) says that a set S contains all causes of E only if S nomically determines E. Suppose that you observe a scenario, Scenario 1, in which A, B, C and D obtain at *t*. At the next instant E obtains. Suppose, furthermore, that you know that {A, B, C, D} includes all of the causes of E that obtain at *t*. However you don't know if *all* members of the set are causes of E or if only some of them are. In particular, you want to test if A is a cause of E. Here's a good way to do it. Compare Scenario 1 to a different scenario, Scenario 2, in which B, C and D obtain, and A does not. Then check to see if E obtains in that scenario. If it *does*, we can't tell anything about whether or not A is a cause of E. After all, it's compatible with set {B, C, D} being nomically sufficient for E that A was in fact a cause. This is just what happens in cases of redundant causation. Suzy's throw was a cause of the window shattering even though Suzy's throw was not necessary to nomically determine that the window shattered.

If E does *not* obtain in Scenario 2, however, we can use the method of elimination and (D/d**) to infer that A is a cause of E. By (D/d**) we know that {A, B, C, D} includes all causes of E only if {A, B, C, D} nomically determines E. We also know that there is at least one set containing all of the causes of E, and that {B, C, D} does *not* include all causes of E (since it does not nomically determine E). From this we can infer that A is needed for the set to nomically determine E. Thus, A is a cause of E.

The method just described involves observing and then comparing two scenarios that are the same in all relevant ways except that A obtains in one but not the other. But it is not required that we observe two actual scenarios. It is sufficient to show that in a relevantly similar possible world with the same laws and where B, C and D obtain without A, E does not obtain. This is how we derive the counterfactual test from nomic determination. We start with a necessary condition (D/d**) for a set to include all causes of E. With the help of the method of elimination we derive a sufficient condition for some particular event, say A, to be a cause of E. Of course, that sufficient condition is familiar: That E counterfactually depends on A is sufficient for A to be a cause of E.

What makes Kment's account particularly interesting and potentially of great importance, in my view, is that it threatens to undermine the motivation for thinking that causation might be reducible to counterfactual dependence. If Kment's picture is right then the role that counterfactuals play in our causal thinking is already accounted for: counterfactual dependence is *evidence* that two events are causally related, and it is so in virtue of its connection to nomic determination. It would be explanatorily superfluous to countenance the existence of a counterfactual analysis of causation, in addition. It is worth our while, then, to see if Kment's account is right.

If Kment is correct that counterfactual dependence is only useful as a heuristic for testing causal claims in virtue of its relationship to the determination idea, then we might expect that we'd be far more inclined to use counterfactuals to support claims about particular events *being* causes, than we would to support claims about particular events *not* being causes. That's because Kment's method gives us only a one-way test. If N is a necessary condition for A, supposing that there is an A, that nothing other than x is N is a sufficient condition for x to be A. It is not necessary that nothing other than x is N for x to be A, however. The fact that none of the other potential suspects were in the garden at the time of the crime is sufficient for the butler to be the murderer. But it is not necessary for the butler to be the murderer that no one other than he was in the garden. If it turns out that others were in the garden at the time of the murder, this in itself tells us nothing about whether or not the butler is guilty.

In fact, Kment considers it to be a significant advantage of his view that his method yields only a one-way test. That's because the counterfactual test for causation only works in one direction. As we know, it is widely held that E counterfactually depending on C is sufficient for C to be a cause of E. But, as we saw before, it is not necessary that E counterfactually depends on C for C to be a cause of E. So it is perfectly fitting, Kment claims, that from (D/d**) we should only be able to derive a sufficient condition involving counterfactual dependence, and not a necessary one.

As it happens, we do regularly use counterfactual reasoning to determine that something is *not* a cause. And we are often justified in doing so. The question is, can Kment's proposal account for this in a satisfactory manner? Before trying to answer this question, let us consider a couple of examples of when we might use counterfactual reasoning as evidence that something is not a cause. Consider Kment's own example of baking a cake. Suppose that you bake a cake and it comes out chalky. You want to know if adding too much flour was the cause. You make the cake again and do everything just as before except this time you add less flour, and this time the cake comes out perfectly. Assuming that all other relevant factors were the same both times, it is reasonable to conclude that the extra flour caused the cake to be chalky the first time. Here's what Kment says.

> In this procedure you are considering a possible situation in which you are not using [extra flour] to make the cake, but which is like the actual situation in all other relevant ways, and which follows the same laws. And you figure out that in such a situation, the cake does not taste chalky. That is a simple version of counterfactual reasoning. Our discussion therefore suggests that the method of evaluating causal claims by counterfactual reasoning is simply an extension of the method of difference. It works in essentially the same way: it relies on the determination idea to establish the causal claim by the method of elimination. (2010: 94-95)

This seems plausible. But now suppose that the second cake you bake turns out just as chalky as the first. Here we have a case where a potential cause in question (the amount of flour used) varies, but the effect (chalky cake) stays the same. It seems just about as reasonable to conclude from this that it *wasn't* the additional flour that caused the chalkiness. There must have been a different cause.

We reason this way all of the time. Here is one more example. A poison kills its victim more slowly when taken on a full stomach. In discussing this case, David Lewis invokes common sense when he denies that eating dinner prior to drinking the poison is a cause of the victim's death. It is plausibly a cause of the death having been slower and more painful than it otherwise would have been, but is not a cause of the death itself (1986:xx). Lewis is surely right here, but we might now ask what grounds this intuition. Here's one very natural reply. The meal wasn't a cause of the death because the death would have occurred in the same basic manner (i.e., via the poison) whether the victim had eaten beforehand or not. That is, the food made no difference to whether or not the victim died from the poison. We reason from the fact that E does not counterfactually depend on C to the conclusion that C is not a cause of E.

Since Kment's main thesis is that what justifies us in using counterfactual reasoning in our causal thinking is (D/d**) and *not* that causation somehow consists in or reduces to counterfactual dependence in some way, it is important for his proposal to be able to account for the value of counterfactual reasoning in the other direction too: that is, from the absence of counterfactual dependence to the absence of a causal connection. How might Kment do this by appeal to (D/d**)?

As we know, if {A, B, C, D} includes all of the causes of E that obtain at *t*, and if, in a scenario in which {B, C and D} (but not A) obtain and E does not, we can use Kment's method to conclude that A is a cause of E. But if E *does* obtain, we cannot conclude anything about whether or not A is a cause. That is, unless for a set to include all causes of E it is not only necessary but also sufficient for that set to nomically determine E. We know that this certainly is not the case in general. But perhaps in those particular circumstances in which it seems reasonable to conclude that C does not cause E based on the fact that E does not counterfactually depend on C, we can justify the inference like this. By (D/d**) it is necessary for the set of E's causes to jointly nomically determine E. If E occurs in the scenario in which B, C and D obtain but A does not, then, still assuming determinism, B, C and D nomically determine E by themselves – that is, without A. How do we get from this to the conclusion that A is not a cause of E? If we also know (or have sufficiently good reason to believe) that there is only one causal pathway to E – i.e., that there is only one set of particular facts minimally sufficient to nomically determine E – then we can conclude with corresponding confidence that A is not part of the only minimally sufficient set and thus, that A is not a cause of E.

There are two separate types of claims one could make in support of Kment's view. The first is that something similar to the method above is how we actually reason when we infer the absence of causation from the absence of counterfactual dependence. This is exceedingly implausible. We make this inference easily and naturally, and just a bit of introspection can tell us that we do not run through anything like the elaborate and complicated steps described above each time. The second, much better possibility is that (D/d**), in conjunction with the oftentimes reasonable-seeming assumption that there is only one set of actual events which are minimally sufficient to nomically determine E, is what justifies our appeal to the absence of counterfactual dependence as evidence for a lack of causal connection, even if we do not generally explicitly reason in this manner.

While this second possibility is certainly more plausible, it still seems rather implausible given the ease and naturalness with which we use counterfactuals to draw causal conclusions (in both directions). I do not deny that we are capable of reasoning using "short-cuts" or heuristic devices – for all I know this could be something we do frequently. However, there should be some way to explain how we develop confidence that the reasoning shortcut or heuristic device (usually) works. Perhaps we recognize what it is a shortcut for, and so can see that we are justified in using the shortcut in virtue of our justification for using the longer method. Or maybe others have figured out that there is a reasoning shortcut and have explicitly taught it to us (e.g. as when a math student uses the quadratic equation without necessarily having any understanding of why it works). Or we have figured out that the shortcut works by induction. That is, we have experience using the shortcut and observing that it is effective. For instance, I may get used to making an inference from the sky appearing a certain way to the likelihood that there is a storm approaching, even if I don't understand the connection.

None of these seem to apply in this case, though. The first two can be ruled out fairly quickly, I think.[22] I contend that the third can be ruled out as well. Induction can work when one receives feedback on the success of the method. I learn that there is a connection between a particular appearance of the sky and an approaching thunderstorm because I observe that one frequently follows the other. But we don't get the same kind of feedback regarding the veracity of our causal inferences. I don't get to observe whether or not my intuition is right that if the death by poison would have still happened had the victim not eaten a meal first, then the meal is not a cause of his death. I don't get feedback on the strength of my inference that if I had used less flour (and done everything else exactly the same) and the cake had still come out chalky, then the flour I removed was not a cause of the chalkiness. At the very least if we are to take seriously the proposal that the counterfactual reasoning we use to make and defend our causal inferences is itself only

[22] An alternative possibility is that we just got "lucky": all this time we've been using counterfactuals to make causal inferences and as it turns out, because of (D/d**) we were right to do so.

justified through an elaborate series of steps involving (D/d**), we are owed a story for why reasoning counterfactually about causes and their absences comes so very naturally and easily to us (almost as though counterfactual dependence is an integral part of our very concept of a cause).

There is an additional, more serious problem for Kment. Consider the case of the soldier, the sergeant and the major.

Soldier, Sergeant and Major The soldier is ready to obey any and all orders made by his superiors, among whom are the sergeant and the major. Because the major is ranked more highly than the sergeant, if the two officers each give an order at the same time, the soldier will obey the major rather than the sergeant. As it happens, both yell out “advance!” simultaneously, and the soldier advances.[23]

Whose order caused the soldier to advance? Intuitively it was the major’s. But why should this be so? The soldier’s advancing does not counterfactually depend on either order. Had the major stayed silent, the soldier would still have advanced because of the sergeant’s order. Likewise, had the sergeant stayed silent, the soldier would still have advanced because of the major’s order. The two cases are not entirely parallel though. There is one disanalogy. That is that the soldier follows the major’s orders in all nearest worlds in which both make (convergent *or* divergent) commands; it just so happens that in the actual world they make the same command. In other words, had the major and the sergeant simultaneously shouted conflicting commands instead of the same command, the soldier would have obeyed the major. This, of course, is a counterfactual, and it is a counterfactual that seems to be intimately connected to the fact that it is the major’s order, and not the sergeant’s, that is the cause of the soldier advancing.

[23] Jonathan Schaffer (2000) attributes this case to Bas van Fraassen.

But if that is right, it indicates that the relationship between difference-making and causation is not nearly as straightforward as Kment's view suggests. The truth of counterfactual M: <if the major had commanded the soldier to do something else, the soldier would have done what the major commanded> indicates that there is a certain kind of dependence between what the major commands and the soldier's actions. The soldier's actions vary with the major's commands in certain relevant circumstances. But it is not clear that it is a kind of dependence that can be accounted for using Kment's method.

Like in the cases previously discussed, we're using a counterfactual as evidence that something is (and something else is not) a cause of the effect in question. And yet, we can't use Kment's method to account for the value of our counterfactual reasoning in this case. That's because we are *not* using a counterfactual scenario to compare a set of potential causes with a set that includes all but one of these to determine if the effect still obtains. Here, rather, we use the counterfactual scenario to compare the actual world where some set of potential causes {A, B, C, D} obtain with nearby worlds where one of those particular facts, say A, is replaced with another fact, rather than just eliminated. So we are comparing the actual scenario where the major said "advance!" at t not with worlds in which he merely didn't command "advance!", but in which he commanded something else. And we are then checking to see if the effect (the soldier's advancing) obtains when the fact is replaced by another. How would Kment's method work, here? Let's try applying it to this case.

Suppose that you observe a scenario (Scenario 1) in which A, B, C and D are present at time t. At the next instant E obtains. Suppose, furthermore, that you know that the set {A, B, C, D} includes all of the causes of E that obtain at t, but you don't know if *all* members are causes of E or only some of them. In particular, you want to test if A is a cause of E. So you compare Scenario 1 to a different scenario, Scenario 2, in which B, C and D obtain, and F obtains instead of A. You then check to see if E obtains in that scenario. Suppose that E does *not* obtain in Scenario 2. What can be concluded? By (D/d**) we know that there is at least one set containing all of the causes of E. We also know that {B, C, D, F} does *not* include all causes of E (since it does not nomically determine E). By (D/d**) we know that {A, B, C, D} includes all causes of E only if {A, B, C, D} nomically determines E. We also know that there is at least one set containing all of the causes of E.

What can we infer from this? Can we infer that A is a cause of E? We cannot. We can infer only that *either* A is a cause of E or else that F *undermines* the obtainment of E. That is, it is compatible with the above that B, C and D jointly nomically entail E given the absence of F, but that the obtainment of F makes E not obtain. Consider the cake example again. If in Scenario 1 we bake a cake which comes out chalky, and if in Scenario 2 we bake a cake in the same way as before except that we replace some of the flour with butter and it doesn't come out chalky, we can conclude either that the extra flour caused the chalkiness *or* that the butter made the cake less chalky, and perhaps would have done so even if the same amount of flour had been used as in Scenario 1.

In Soldier, Sergeant and Major, if we use Kment's method the most we can infer from the counterfactual scenario in which the major makes a different command and the soldier doesn't advance is that either the major's order caused the soldier to advance *or* the major's alternative command caused the soldier to *not* advance in the counterfactual scenario (or both). This could still get us to the conclusion that the major's order caused the soldier to advance in the actual world if we were able to somehow rule out the second disjunct. But in fact, the second disjunct seems *true*: the major's alternative command in the second scenario *is* a cause of the soldier not advancing (since he would otherwise advance due to the sergeant's command to advance). So if it's right that counterfactual M is tied to it being the major's order that causes the soldier to advance, it cannot be for the reason that Kment's account suggests. But if that's right, counterfactual reasoning and difference-making are related to causation in some other way: not merely in virtue of (D/d**).

II. Sartorio (2005) offers an altogether different kind of suggestion for how difference-making and causation are related. Sartorio begins by arguing that David Lewis's influential (1986) treatment of difference-making ends up counting too many things as causes. Here's Lewis.

> Lewis: C causes E if and only if there is a chain of stepwise counterfactual dependence from C to E (1986, p. 167)

To show that Lewis's proposal erroneously rules things as causes which are not in fact causes, Sartorio asks us to consider Switch, shown below.

> *Switch*: Victim is stuck on the railroad tracks. A runaway train is hurtling
> down the tracks when it approaches a switch. I flip the switch, and the train
> turns onto a side track. However, the tracks reconverge a bit further ahead,
> before the place where Victim is standing. Victim dies. (2005: 73)

If Lewis's proposal above is correct, flipping the switch is a cause of Victim's death. Here's Sartorio's explanation:

> …there is a chain of stepwise counterfactual dependence from my flipping the switch to the death, via the intermediate event of the train running on the side track. This emerges as follows: The train's running on the

> side track counterfactually depends on my flipping the switch, for, had I not flipped the switch, the train wouldn't have run on the side track. In turn, given that the train switched tracks and thus it is no longer on the main track, Victim's death counterfactually depends on the train's running on the side track. For, if it hadn't been running on the side track, then, given that it is not running on the main track, the train would not have reached Victim and killed him. Hence, the train's running on the side track counterfactually depends on my flipping the switch, and Victim's death counterfactually depends on the train's running on t he side track. So [Lewis's proposal] entails that my flipping the switch caused Victim's death in Switch. (2005: 73)

But intuitively, the flipping of the switch does not make the relevant kind of difference to Victim's death to be a cause of his death. Sartorio observes that not only does the flip not make a difference to the death, but a failure to flip the switch wouldn't have made a difference to the death, either. This motivates the idea that "the reason that my flipping the switch doesn't make a difference is that the contribution that it makes is not more important than the contribution that its absence would have made. Maybe, for something to be a cause, it must make a contribution that somehow outweighs the contribution that its absence would have made." (2005: 75) She proposes the following "Causes as Difference-Makers" principle:

> CDM: If C caused E, then, had C not occurred, the absence of C wouldn't have caused E.

CDM is intended as a constraint on an analysis of causation. It captures the intuitive idea that a genuine cause contributes more to the effect than its absence would have. According to Sartorio, CDM rules that the flip is *not* a cause in Switch, since the flip and the failure to flip intuitively make the same kind of contribution to the death, and so if the flip were a cause its absence should count as a cause as well. But this violates CDM.

I don't think that CDM is the right principle to capture the kind of difference-making relevant to causation, and seeing why will prove helpful for my investigation into difference making, here. Consider the following case.

Soldier and Captain A soldier is given the following instructions. If at *t* the captain either tells the soldier to fire or doesn't say anything at all, the soldier must fire. If at *t* the captain says "hold your fire" the soldier must abstain from firing. The captain says "fire" and the soldier fires.

It seems that the captain's command was a cause of the soldier's action. After all, the captain could have instead said "hold your fire" and the soldier would not have fired. It also seems that if the captain had stayed silent at *t* and the soldier had fired, the captain's silence would be a cause of the soldier's firing. After all, rather than staying silent the captain could have commanded the soldier to hold his fire. But by CDM the captain's saying "fire" cannot be a cause of the firing if the absence of the command also causes the soldier to fire.

Both Kment's and Sartorio's proposals implicitly make the commonly-held assumption that when we think about difference-making and its relationship to causation we can think of the possible states of events in binary terms: either a given event obtains in the actual world or it doesn't. Either the given event obtains in the relevant counterfactual scenario or it doesn't. In Soldier, Major, Sergeant the relevant type of difference-making seems to be this: varying *what* the major commands (rather than *whether* the major commands) makes a difference to which action the soldier performs (and not just to whether or not the soldier advances). Soldier and Captain is a bit different. Here the actions of the soldier depend not only on whether or not the captain orders "fire!", but also on what the captain does if he doesn't command the soldier to fire. Does he stay silent or does he command something else? Both Soldier, Major and Sergeant and Soldier and Captain suggest that we need to broaden our notion of difference-making. In future work I intend to show how doing so will get us closer to a viable counterfactual analysis of causation.

REVERSE SOBEL SEQUENCES AND WHY MOST UNQUALIFIED 'WOULD'– COUNTERFACTUALS ARE NOT TRUE

I. Introduction

Sophie is considering going to the 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

But 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.

Sequences like (1), commonly referred to as *Sobel sequences*[24], famously motivated David Lewis (1973) to reject the once popular strict conditional semantics of counterfactuals in favor of the now classic *variably* strict conditional semantics. The strict conditional analysis treats a counterfactual as a material conditional embedded under a necessity operator, where the domain of worlds in the operator's scope is determined by the context. In symbols, if '>' represents the counterfactual connective, '⊃' the material conditional connective and '□' the necessity operator, the strict conditional account says that for any two propositions ϕ and ψ, $\phi > \psi =_{\text{def}} \Box (\phi \supset \psi)$. Call the worlds in the necessity operator's domain (in a context) the *accessible worlds*.

What happens when we evaluate (1a) and (1b) using the strict conditional semantics? If (1a) is (nonvacuously) true, then at all of the accessible worlds at which Sophie goes to the parade,

[24] Lewis (1973) thanks J. Howard Sobel for bringing sequences like (1) to his attention.

she sees Pedro. If that is so then Sophie sees Pedro at all of the accessible worlds at which she goes to the parade and is stuck behind someone tall. But in that case, (1b) is false (unless there are no accessible worlds where Sophie goes to the parade and is stuck behind someone tall – in which case (1b) is vacuously true). The problem for the strict conditional analysis is that it entails that (1a) and (1b) cannot both be nonvacuously true in a fixed context.[25]

Following Robert Stalnaker (1968), Lewis's (1973) solution is to order the accessible worlds according to their similarity to the world of assessment (*w*), based on a particular, contextually-determined similarity metric. Pictorially we can imagine this, as Lewis does, by picturing a system of spheres centered around *w*. Each sphere represents a degree of similarity (or 'closeness') to *w*. The worlds represented in the innermost sphere are the worlds most similar to *w*. Moving outward, worlds represented in each successively larger sphere are worlds that are successively less similar to *w*. The worlds represented in the region outside the system of spheres are the inaccessible worlds. For Lewis (but using my notation) a counterfactual A>C is non-vacuously true just in case there is an 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.[26]

According to Lewis's variably strict conditional semantics described above, if all of the Sophie-goes-to-the-parade-worlds most similar to the actual world are worlds where Sophie sees

[25] Lewis considers the possibility that counterfactuals could be "vague strict conditionals based on similarity, and that vagueness is resolved – the strictness is fixed – by very local context: the antecedent itself." But he rejects this, saying that it "…is not altogether wrong, but it is defeatist. It consigns to the wastebasket of contextually resolved vagueness something much more amenable to systematic analysis than most of the rest of the mess in that wastebasket." (1973: 13) As we will see, von Fintel (2001) and Gillies (2007) endorse a picture very similar to the one Lewis is rejecting here.

[26] 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.

Pedro, then (1a) is true.[27] But the truth of (1a) does not preclude the truth of (1b). (1b) is also true if the nearest worlds at which Sophie goes to the parade *and* is stuck behind someone tall are worlds where Sophie does not see Pedro. As long as all of the worlds at which Sophie goes to the parade but is stuck behind someone tall are less similar than the most similar worlds where she goes to the parade and sees Pedro, both counterfactuals can be nonvacuously true.

So far so good. Unlike the strict conditional semantics, the variably strict conditional semantics handles Sobel sequences very well. But there is a problem. The problem of *reverse Sobel sequences*, attributed to Irene Heim and first discussed by Kai von Fintel (2001) and Anthony Gillies (2007), is that if the order of (1a) and (1b) is reversed, it no longer seems that both counterfactuals are true:

(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.[28]

Given that both (1a) and (1b) seem true, the infelicity of (2b) is unexpected if the variably strict conditional analysis is correct. If the closest worlds where Sophie goes to the parade are worlds where she sees Pedro (as (1a) says), and if the closest worlds where she goes to the parade and is stuck behind someone tall are worlds where she does not see Pedro (as (1b) says), then, according to the variably strict conditional semantics, the counterfactuals in sequence (2) should both be true, regardless of which is uttered first.

[27] For ease of exposition I will generally speak as though the 'Limit Assumption', the assumption that there is always a set of most similar antecedent-worlds, holds. Lewis (1973) denies the Limit Assumption but nothing hangs on that here.

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

This challenge for the classic model has been taken very seriously in the literature on counterfactuals. It has led some theorists (von Fintel (2001), Gillies (2007)) to reject the variably strict conditional semantics entirely, and instead endorse a variation of the original strict conditional account. Others (Ichikawa (2011), Karen Lewis (2017)) have argued 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 argues that, in fact, the classic semantics can handle reverse Sobel 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*.

It is my contention that none of these reactions to the problem of reverse Sobel sequences is the right reaction. After showing why I think each of the proposals is inadequate, I will defend a novel way to make sense of the troublesome sequences. The solution I endorse avoids the problems faced by the alternative analyses. In addition, there is good independent reason to think that it is right. There is, however, a difficulty for my view: its truth entails that many ordinarily accepted counterfactuals are *not* true. I will argue that this (apparent) cost is an acceptable one. Before defending my own solution to the problem, however, we should take a brief look at the three extant proposals.

II. Von Fintel and Gillies: the Dynamic Semantic Solution

In response to the reverse Sobel sequence problem, von Fintel (2001) and Gillies (2007) have each(independently) argued that Lewis and Stalnaker were wrong to reject a strict conditional analysis in favor of the variably strict conditional analysis in the first place. Instead, they contend, the strict conditional semantics is basically correct – it just needs some tweaking. 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' demands that there are accessible A-worlds in the domain.[29] 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 where Sophie goes to the parade and gets stuck behind someone tall. But if that is so then (2b) is false: it is not the case that Sophie sees Pedro in 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 be accessible worlds where 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 where Sophie goes to the parade. Accommodating the (weaker) demands of (1b) requires that we bring into the domain worlds where Sophie gets stuck behind someone tall, and at these worlds, Sophie does not see Pedro. Thus, on this picture, (2a) and (2b) are semantically inconsistent, though (1a) and (1b) are not. This accounts for the felicity of sequence (1) and the infelicity of sequence (2). As we will see in the next section, however, there are good reasons to reject the dynamic semantic account.

III. Moss: the Pragmatic Solution

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

Moss (2012) has defended an entirely pragmatic way to account for the infelicity of reverse Sobel sequences. And, as she has shown, her solution is superior to the solution of Von Fintel and Gillies in at least two important ways. On Moss's view, (2b) is *not* infelicitous because it is *false*, as on the dynamic semantic analysis. Rather, (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 is such that (i) it cannot be ruled out by the speaker and (ii) it is incompatible with the assertion, the assertion is infelicitous. Since the speaker cannot rule out that if Sophie had gone to the parade she would have been stuck behind someone tall and so not seen Pedro, the possibility, when raised to salience by (2a), makes (2b) pragmatically infelicitous. (According to Moss (1a) does not similarly make (1b) infelicitous because that Sophie would have seen Pedro (had she gone to the parade) is intuitively no longer part of the common ground once (1b) has been uttered.[30])

Pragmatic infelicity of the kind Moss describes is a widely occurring phenomenon instantiated by a wide variety of different kinds of utterances, not just counterfactual ones. Here is one of Moss's examples. Suppose you and I are looking at the zebras at the zoo, and we have the following exchange:

(7) a. That animal was born with stripes.
b. But cleverly disguised mules are not born with stripes.

As Moss writes, "[t]his reply may be a non sequitur, perhaps even a little annoying. But otherwise there is nothing wrong with [the] reply." (2012: 567) Things change when we reverse the order of

[30] Moss claims that this asymmetry (between (1) and (2)) in whether the first proposition remains common ground once the second sentence of the sequence has been uttered is a ubiquitous kind of asymmetry. See Moss (2012: 570-571)

(a) and (b), however:[31]

(8) a. Cleverly disguised mules are not born with stripes
b. #But that animal was born with stripes.

Moss concludes,

> So why is [(8)] bad, while [(7)] is okay? Here is one intuitive answer: in the above scenario, [(8b)] is infelicitous because [(8a)] raises the possibility that the caged animal is a cleverly disguised mule, and the speaker of [(8b)] cannot rule out this possibility. So [(8b)] is infelicitous because in the above scenario, it is an epistemically irresponsible thing to say. (2012: 568)

Moss gives two important considerations that rule in favor of her pragmatic explanation of the infelicity of reverse Sobel sequences over the semantic explanation given by von Fintel and Gillies. The first is that, as we have seen, her explanation is far more general. It accounts for a much wider range of data.

The second consideration is that Moss's solution can provide a natural explanation for why in some cases the second proposition of a reverse Sobel sequence does *not* sound infelicitous. And there are a lot of good examples of sequences that are like this. Moss gives the following helpful example (2012: 574):

> Suppose John and Mary are our mutual friends. John was going to ask Mary to marry him, but chickened out at the last minute. I know Mary much better than you do, and you ask me whether Mary might have said yes if John had proposed. I tell you that I swore to Mary that I would never actually tell anyone that information, which means that strictly speaking, I cannot answer your question. But I say that I will go so far as to tell you two facts:

(9) a. If John had proposed to Mary and she had said yes, he would have been really happy.
b. But if John had proposed, he would have been really unhappy.

[31] There are ways to read (8) such that it sounds perfectly fine. The person uttering (8b) has the option to resist taking seriously the possibility made salient in (8a), i.e., that the animal could be a disguised mule with stripes (maybe the possibility seems too outrageous). As we will see shortly, Moss can account for this as well.

Here the second counterfactual in the reverse Sobel sequence is felicitous, and Moss can explain why. The speaker can rule out that Mary would have said yes, had John asked (of course, this is exactly what the speaker is intending to communicate by asserting the two counterfactuals). And if the speaker can rule out that Mary would have said yes, then the case does not meet Moss's first criterion, stated above, for instantiating the kind of pragmatic infelicity instantiated by sequences like (2) and (8).

Moss's pragmatic explanation of the infelicity is very appealing. There are, however, some problems with it. Moss takes herself to be defending Lewis's and Stalnaker's original variably strict semantics: if her solution is right, von Fintel and Gillies were wrong to think that reverse Sobel sequences pose a problem for the classic theory. But let us consider what it would mean on the classic model for (2b) (<If Sophie had gone to the parade she would have seen Pedro>) to be true even though the speaker cannot rule out the possibility that Sophie could have gone to the parade and been stuck behind someone tall. There are special problems if it is Lewis's semantics that is the right one. As we have seen, on the standard Lewisian picture a counterfactual is true just in case there is an A&C–world closer to the actual world than any A&Not-C – world. Consider sequence (2) again:

(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

If (2a) raises to salience a possibility that is incompatible with (2b), it must raise to-salience the possibility that *not* all nearest A–worlds are C–worlds: in other words, that at least one of the nearest A–worlds is a not–C–world.[32] Now if it were *actually* the case that at least one of the closest Sophie-goes-to-the-parade worlds is a world where Sophie gets stuck behind someone tall (and so does not see Pedro), (2b) would be false. But recall that for Moss it is a crucial part of her view that (2b) can be true. For (2b) to be true on the classic model it needs to be the case that there are not any A&not–C worlds as similar as the nearest A&C world. Thus, given Moss's account, if (2b) is true despite being infelicitous, its infelicity must be because those assessing (2b) do not *know* that (2b) is true (if they could know it was true, presumably it wouldn't be infelicitous). That is, those hearing (2b) as infelicitous must take it as a legitimate possibility that there is an A&not–C world at least as close as the nearest A&C world, despite it being the case that in fact, all the nearest A—worlds are C—worlds.

This is not necessarily a problem in itself. Perhaps we should expect people to be *consistently* in doubt about whether all the nearest A-worlds are C-worlds, even in the cases where in fact, all nearest A-worlds *are* C-worlds. (I say "consistently" because a reverse Sobel sequence like (2) *always* sounds infelicitous unless some very particular kind of background story is given to counteract this.) The more serious worry is that unless Moss is happy to deny (with Stalnaker and contra Lewis) that *would-* and *might-*counterfactuals are duals, it seems that she must be committed to the infelicity of (1c):

(1) a. If Sophie had gone to the parade she would have seen Pedro.
b. But if Sophie had gone to the parade and been stuck behind someone tall, she would not have seen Pedro.
c. $_{\text{ok}}$ I suppose you're right. If Sophie had gone to the parade she might not have seen Pedro.

[32] Karen Lewis (2017) is concerned with identifying exactly what possibility is raised to salience, on the best version of Moss's view. She ultimately concludes that there is no good option and uses this as a reason to reject Moss's account.

Let me explain. If would- and might-counterfactuals are duals (i.e., ϕ–might–ψ is equivalent to <not (ϕ–would–not–ψ)>), then (1c) is inconsistent with (1a). (1a) says that all closest A–worlds are C–worlds and (still assuming would- might- counterfactual duality) (1c) says that it is not the case that all closest A–worlds are C–worlds. If Moss's account and the duality thesis are both right then we should expect (1c) to sound infelicitous. That is because the possibility that if Sophie had gone to the parade she would have seen Pedro (i.e, (1a)) is clearly salient in the context, unless (1b) has the power to undermine it. But the possibility raised to salience by (1b) does not, on Moss's view, entail that (1a) is false. (1a) can still be true, on Moss's view, even if expressed *after* (1b) (that is after all what is going on in sequence (2)). So Moss is committed to the truth of (1a) still being possible, even after (1b) has been uttered. But if (1a) can still be true (i.e., it cannot be ruled out), and if (1a) is inconsistent with (1c) (which is the case given the duality thesis), we should expect, given Moss's view, for (1c) to be infelicitous. And yet it is not infelicitous. In fact, (1c) is exactly what we might expect someone to say following the utterance of (1b).

Karen Lewis (2017) has independently noticed that the inconsistency of (1a) and (1c) poses a problem for Moss if the duality thesis is correct, but draws a conclusion that I think is weaker than what is justified. From the inconsistency of (1a) and (1c) Karen Lewis concludes that,

> [I]f one wants to use the [Moss]-style explanation to preserve the classic Lewis semantics, one cannot call such sequences consistent, as the dynamic account can, and one is committed to the existence of many false propositions that are perfectly assertable [(since, if by hypothesis (1a) and (1b) are both true, (1c) must, on David Lewis's view, be false).] ...being committed to the existence of a plethora of assertable false propositions is a high cost, especially for anyone who wants to maintain the knowledge norm of assertion. There's the further problem that while [(1a)] seems perfectly true in its context, [(1c)] also seems perfectly true in its context of utterance. Moss can at most account for the felicity, not the truth of [(1c)], at least not without bringing in additional machinery to the basic Lewis semantics. (2017: 18-19)

But in fact, the problem is more serious than this. It is not merely that Moss must be committed to the existence of a plethora of assertable false propositions. Nor is the problem merely that Moss cannot account for (1c)'s apparent truth. As we have seen, Moss cannot account for the *felicity* of (1c), either. To repeat very briefly, this is because Moss claims that an utterance will be infelicitous if there is a salient possibility that the speaker cannot rule out, and which contradicts the original assertion. And there is a salient possibility that contradicts (1c): namely, the possibility expressed by (1a), that all nearest A-worlds are C-worlds.

Of course one way out of this difficulty is for Moss and proponents of her view to reject the duality thesis. I suspect that Moss would not herself be happy with this solution given that many theorists accept would- and might-counterfactual duality, and, in addition, Moss has stated her intention to remain neutral about the duality thesis (2012: 571). But even if it turns out that would- and might-counterfactuals are *not* duals, there are additional difficulties. In particular, as we will see in section V, there are reasons to think that Moss's account cannot tell the full story. The most important difference between the dynamic semantic model and Moss's pragmatic solution is that according to the former (2b) is *false*. If Moss is right (2b) may very well be true, just as (8b) <But that animal was born with stripes> may be true, even if asserted after (the true) <Cleverly disguised mules are not born with stripes>. So if it were to turn out that the second counterfactual in a reverse Sobel sequence is *always* false when the sequence is infelicitous, this would tell strongly against Moss's proposal, which lacks the resources to account for why this would be so.[33] While I

[33] Note that if von Fintel and Gillies are right, the second counterfactual of the reverse Sobel is not merely false when the sequence is infelicitous. Unless the defender of the dynamic semantic account can give a reason to think otherwise, it seems that on that view we should expect the second counterfactual to *always* be false. But this makes falsity far too ubiquitous, and indeed, is the basis of another objection to their semantics: Clearly Moss's (9b) <if John

will not argue that the second sentence of an *infelicitous* reverse Sobel sequence is always false, I will argue that it is false in all of the interesting cases. Since Moss's solution cannot account for even this more limited claim, the limited claim's truth would suffice to show that a different explanation is needed. Before getting to my defense of the limited claim, however, we should consider one final proposal.[34]

IV. Karen Lewis: the Contextualist Solution

Karen Lewis ((2015), (2017)) has a novel solution that evades the problems faced by the alternative 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 picture the semantic value of a counterfactual is sensitive to features of the conversational context such as e.g., the standards of precision and the salience (or non-salience) of possibilities. Formally, the difference between Karen Lewis's semantics and the Lewis-Stalnaker semantics is that while on the classic model possible worlds are ordered just according to how similar they are to the world of assessment, for Karen Lewis worlds are ordered by both similarity and *relevance*. "The only thing we need to add to [the Lewis-Stalnaker semantics] for the time being to get the account off the ground", she writes, "is that the similarity ordering or selection function is influenced by relevant salient possibilities, so that the closest worlds as determined by a given context include relevant salient possibilities [and exclude ignored, non-salient possibilities]" (2017: 20). In a bit more detail:

had proposed to Mary, he would have been really unhappy> need not be *false* even if the previously uttered (9a) <If John had proposed to Mary and she had said yes he would have been really happy> is true.

[34] For some additional objections to Moss see Karen Lewis (2017).

> ...the closest worlds are not just the most similar worlds, but...both similarity and relevance contribute to what counts as a closest world: some worlds that are most similar but aren't relevant are not among the closest worlds, and some worlds that are relevant but that are not the most similar are among the closest worlds. Picturesquely, this is how the ordering sources interact: relevance can take worlds that are among the most similar worlds and move them farther away, so that they are not among the closest worlds. It can also take worlds that are less similar – as long as they are similar enough (which is vague) – and move them closer, so that they are among the closest worlds. (2017: 23)

Her truth conditions are given below.

> For all contexts c, P □→ Q is true in c iff all the closest P-worlds are Q-worlds, where closeness is a function of both similarity and relevance. (2017: 22)

But what exactly is relevance? Karen Lewis characterizes it as an *objective* feature of the conversation, although by this she just means that "speakers are limited in how much they can affect what is relevant and irrelevant". Not every possibility can be made relevant if brought to salience (i.e., not any less-similar world can become among the closest) and not every possibility can be made irrelevant by being ignored (i.e., some most-similar worlds will be among the closest no matter what). In other words, there are objective constraints on what must, can, and cannot be relevant in a context. It will be important, for what follows, for us to see what kinds of constraints Lewis has in mind. They include the following. "High probability macroscopically-described outcomes are always relevant" and "...in general, relevance doesn't apply to possibilities that are too dissimilar; it can only order among the closest worlds that are similar enough (where similarity is left purposefully vague)." (2017: 27) In addition, the truth matters. So, for instance, "if conversational participants (perhaps justifiably) think that Sophie is extremely shy and is very likely to cower at the back of the crowd at a parade, if the facts are such that they are wrong – Sophie is not shy in this way – the possibility that she has an unblocked view is relevant. Such possibilities cannot be legitimately ignored." (2017: 25)

Now that we have the basic picture let us see how it gets used to solve the reverse Sobel sequence problem. Take sequences (1) and (2) again:

(1) 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.

(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.

Supposing that (1a), (1b) and (2a) are true, (2b) can still be false if the possibility raised to salience by (2a) – i.e. that 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 where 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 where Sophie is stuck behind someone tall and so does not see Pedro are among the closest Sophie-goes-to-the-parade-worlds, then (2b) is false.

Karen Lewis's analysis avoids many of the difficulties faced by the alternative proposals. For instance, it can easily account for the felicity of felicitous reverse Sobel sequences: if the possibility raised to salience by the first counterfactual of the reverse Sobel is not *relevant*, it does not affect the closeness ordering. Consider Moss's sequence (9) again:

(9) a. If John had proposed to Mary and she had said yes, he would have been really happy.
b. $_{\text{ok}}$ But if John had proposed, he would have been really unhappy.

If the possibility that John proposed to Mary and she said yes is not conversationally relevant despite being made salient by (9a), (9b) can be true.

Furthermore, the truth of would- and might-counterfactual duality does not threaten Karen Lewis's contextualism. The problem for Moss, recall, is that on her view (1a) can be true even if asserted after (the true) (1b). But since the duality thesis entails that (1a) is inconsistent with (1c), if the duality thesis is right, that (1a) could be true should make (1c) infelicitous. But (1c) is felicitous. Karen Lewis's analysis does not have this problem. According to her semantics (1a) is no longer true once the possibility that Sophie went to the parade and was stuck behind someone tall is salient, assuming that it is a relevant possibility. (And it seems that it must be relevant for (1c) to be felicitous – otherwise the appropriate reaction to (1b) would be to dismiss the possibility, not to grant that Sophie might not have seen Pedro.) Since it *is* relevant, (1a) is not true at the time of (1c)'s utterance. So there is nothing problematic about the truth (and felicity) of (1c).

Despite the many advantages of Lewis's contextualist treatment, I think it is wrong. I now advance two arguments against it.[35]

Argument 1: Counterfactuals are more objective than Lewis's semantics would have it.

One immediate challenge for a contextualist semantics like Karen Lewis's is that counterfactuals like <if Sophie had gone to the parade she would have seen Pedro> seem to have determinate, invariant truth values once the "facts have been fixed", and the similarity ordering has been established.[36] It seems clear that the truthvalue of a statement like <Agnes is rich> or <Lester is tall> depends on the standards for richness, or for tallness in the context. But

[35] For some additional arguments against contextualism about counterfactuals in general (i.e., not against Karen Lewis's contextualism in particular), see Hájek (ms).

[36] Of course, most agree that counterfactuals are context-sensitive in one important sense: which facts count toward the world similarity ordering depends upon the context. Karen 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." (2015: 16). In other words, there is additional work for the context to do once the similarity function is (contextually) determined.

counterfactuals like (1a) seem importantly different. Intuitively, whether or not Sophie would have seen Pedro (had she gone to the parade) is either determined by facts about the world or it isn't. Either way, this seems to be an objective question about the world (although once again, which facts about the world are relevant can be context-dependent). Its answer intuitively does *not* depend on conversationally-relative facts like which possibilities happen to be salient in the context. And if that's right Karen Lewis's contextualism is wrong.

Karen Lewis aims to preempt a related objection by arguing that the context-sensitivity of the sort she claims counterfactuals exhibit is not in general the kind of thing that undermines objectivity:

> To see that being objective and being dependent on conversational purposes (and thus sensitive to conversational moves) are compatible, compare this view of counterfactuals with one in which the semantics of absolute gradable adjectives like *flat* are sensitive to the context. In a context in which we are looking to play a fair game of pool, I may truly say of a particular pool table, "the table is flat". In a context in which we are looking for a surface for a physics experiment, or where it has been made salient that the pool table has ever so slight bumps on its surface, it is true to say of the same pool table "the table is not flat". This does not make the flatness of the table any less objective, any less sensitive to what the world is actually like. (2017: 26)

While Lewis is clearly right that context-sensitivity and objectivity (of the sort she has in mind) are compatible, the truth-values of assertions involving terms like 'flat' or 'tall' still seem to depend on context in a way that the truth-values of most ordinary counterfactuals do not. In fact, Karen Lewis's analogy with these relatively uncontroversial, paradigmatic context-sensitive expressions is helpful because it makes it easier to pinpoint exactly why counterfactual context-sensitivity of the sort she defends *is* problematic in a way that, for instance, the context-sensitivity of absolute gradable adjectives is not. When we say of a pool table that it is flat, or of a fifth grade child that he is tall, we are in general saying something about the table, or the fifth grader, relative to some standard or comparative class. If it is objectively true that the table is flat (or the fifth grader 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 domain or comparative class she has in mind when pressed.

But a counterfactual utterance is different. When I assert that <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) a. He is tall!
b. [#?] No he is not. Remember the NBA players we saw at the game last night?
c. $_{\text{Ok}}$ Stop being a smart alec, you know that's not what I meant. I meant that he is tall fora boy his age.

(11) a. It is raining outside
b. # No it is not, the sky is perfectly clear
c. $_{\text{Ok}}$ Stop being a smart alec, you know that's not what I meant. I meant that it is raining where *I* am, in Tucson.

(12) 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 that for our purposes it is flat. It is flat enough to lay your drink on.

(13) a. All the beer is warm
b. [#?] Well, not *all* the beer. Surely someone, somewhere has some chilled beer.
c. $_{\text{ok}}$ That's not what I meant. I meant that all the beer in the house is warm.

(14) a. That man is rich.
b. [#?] No he is not. Have you been to my neighborhood?
c. $_{\text{Ok}}$ Okay, relative to my standards he is rich. Relative to your standards he is not.

Compare:

(15) 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 we plug into '___' some standard (or domain).]

The problematic replies in (15c) are good evidence that the speaker who asserts (15a) 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.

In her (2015) Lewis anticipates and responds to an objection that is similar to, although importantly different from, the objection I am advancing here. According to the objection she discusses, "A contextualist theory…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 anaphora is being used) is not one in which the same words express the truth." (2005: 18, her emphasis) For example, the following discourse is (allegedly) felicitous:[37]

[(16)] 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.

c. 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. (2015: 19)

And yet, the objection continues, analogous discourses involving counterfactuals, like (17) below, are *in*felicitous.

[37] I say it is *allegedly* felicitous because to my ear the first sentence of (16c) is infelicitous.

[(17)] 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. (2015: 19)

Lewis replies to this objection by arguing that in fact a contextualist theory should not be thought to predict that discourses like (17) will be felicitous, since there are examples of discourses which involve relatively uncontroversial context-sensitive expressions and which are infelicitous just as (17) is. For instance:

[(18)] a. That field is flat
b. But the field has many small divets and bumps. So it's not flat.
c. # I didn't say it was. All I was saying is that it was flat for a football field.

[(19)] 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. (2015: 20)

By using the examples Lewis uses, she gives this sort of objection short shrift. Discourses (16)-(19) are not good examples for illustrating what speakers should be able to do if there is context-sensitivity. (18) is a bad example because it is plausible that for *any* given expression, x, it is usually infelicitous to utter "I didn't say 'x'", after uttering 'x', even if the content of 'x' varies by context. This seems to hold even for paradigm context sensitive expressions like weather predicates. Discourse (20), for example, is difficult to make sense of:

(20) It is raining. Oh wait, it just stopped. #But I didn't say it was.

A good explanation for the infelicity is that (as Lewis also notes in her discussion of discourse (18)) the 'that' in 'I didn't say that' can be naturally taken to refer to the sentence (i.e., chain of words) uttered, rather than to the proposition expressed. In (18a) the speaker *did* say (the words) 'the field is flat', and so the first sentence of (18c) comes across as blatantly false.

I suggest that the infelicity of (17) and (19) can be attributed to the fact that 'what I said before is still true' can be naturally interpreted to mean that the proposition that would be expressed, if the sentence were asserted in the present context, is true. Consider (21):

(21) It is raining. Oh wait, it just stopped. #But what I said before is still true.

If (21) sounds bad to you (as it does to me), this is presumably because 'what I said before is still true' can here be naturally understood to mean that it is still raining. To express instead that <it is still the case that when I spoke earlier I was speaking truthfully>, the speaker might instead say 'it is still the case that what I said before *was* true'.

That the infelicity of discourses (17)-(21) can be attributed to superficial problems of word choice becomes apparent when it is observed that if the infelicitous remarks are changed slightly, they are no longer infelicitous. We just saw how to change (21) to make it felicitous. The fact that the speaker can felicitously say 'what I said before was true' still provides evidence that 'it is raining' is context sensitive: although 'it is raining', stated now, is false, the sentence was true when asserted in a different context moments before. We can easily revise (18) to make it felicitous as well. In reply to (18b) the original speaker can retort: 'All I *meant* was that the field is flat for a football field (and *that* is still true)'. Now the speaker is clearly referring to the proposition she had intended to express, and not to the words she had used. And, unlike (18c), this reply is felicitous.

The real objection to Karen Lewis, then, is that we cannot similarly 'fix' the parallel replies in discourses involving counterfactuals. This was already illustrated with discourse (15). It is clearly not the case that the speaker asserting (15a) *meant* to say that Sophie would have seen Pedro relative to some standard. Nor can the speaker felicitously say that it is still the case that what she said before, i.e., <if Sophie had gone to the parade she would have seen Pedro>, *was* true when she said it, assuming that she does not believe the sentence would be true if uttered now.

The lesson is that there are two critical, closely related differences between legitimate context-sensitive expressions and counterfactuals. When using the former speakers can usually (i) make explicit (or at least, approximately point to) the standards or domain they have in mind and (ii) say, of a sentence now false, that it *was* true when uttered in a different context. That speakers cannot felicitously do either of these things when using counterfactuals is good evidence that counterfactuals are not context sensitive in the way that Karen Lewis contends.

<u>Argument 2</u>: Karen Lewis draws the line in the wrong place.

Compare these two sequences:

(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.

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

(22) a. If this match had been wet and struck it would not have lit
b. $_{\text{Ok}}$Okay, but if this match had been struck it would have lit.

Why is (2b) infelicitous but not (22b)? If Karen Lewis's semantics is right, (22b) is felicitous because the worlds at which the match is wet are too far away to be relevant, even when brought to salience. Here is what she says about this very case:

> ...when we are considering a situation in which we are predicting what would have happened if I had a struck a reliable dry match, provided that there are no nearby ways in which the match could get wet (I am not holding it right above a bowl of water, no one is pointing a hose at me, etc.), then worlds in which the match is soaked, even if made salient, are too dissimilar to be relevant for the evaluation of subsequent counterfactuals about what would have happened if I had struck the match. (2017: 29)

But what explains why worlds at which the match is wet must be too dissimilar to become relevant? Of course these worlds are not as similar as worlds at which the match is dry since the match is a dry match. But *Karen Lewis does not hold that only the most similar antecedent worlds count as relevant*: to the contrary, she must deny this, since on her view closeness is a function of both similarity and relevance, and less similar worlds can be included among the closest worlds just because of their relevance. Surely we can devise a scenario where the wet-match worlds are not so far away but where (22b) is still assertable. What if there are a lot of wet matches in the same room? Or what if this match was very close to becoming wet just moments ago (but prior to the moments immediately before the time that the match would have been struck) and is now safely dry? Would that change anything about how we'd evaluate sequence (22)? It seems to me that it would not. If we are referring to a particular dry match, it makes no difference how far away the worlds at which the match is wet (assuming they are not as similar as the dry-match-worlds). Unless there is a chance that the match in question would have become wet if it were going to be struck (in which case (22b) would be *infelicitous*) it seems that in any scenario, no matter how similar the wet-match world, it remains the case the particular (dry) match referred to in (22b) would have lit if struck.

The point is that the best explanation for the felicity of (22) is *not* that worlds where the match is wet and struck are vastly dissimilar from the actual world. Nor does there seem to be a good candidate for a different constraint one could impose on the relevance function to rule out match-is-wet-and-struck-worlds.[38] None of the alternative constraints Karen Lewis discusses can do

[38] A seemingly natural way to get around this problem is to revise the relevance function so that it is only able to *remove* (non-relevant) worlds from the set of closest worlds, and unable to bring less similar (but relevant) worlds into the set

the job. So how do we rule out match-is-wet-and-struck worlds? How about the old fashioned way: the nearest worlds where the match is struck and wet are simply *less* close than the nearest worlds where the match is struck and dry.

V. Deflating the Problem

The kinds of difficulties faced by the proposals on offer suggest a different solution to the puzzle. Compare our original Sobel and reverse Sobel sequences (sequences (1) and (2), respectively) with the sequences about the match:

(23) a. If this match had been struck it would have lit
b. But if this match had been wet and struck it would not have lit.

(22) a. If this match had been wet and struck it would not have lit
b. $_{\text{Ok}}$ Yes, but it's still the case that if this match had been struck it would have lit.

We just saw that we cannot account for the felicity of (22b) by insisting that the worlds where the match is wet and struck are extremely dissimilar from the actual world: (22b) is still felicitous when we fill in background details that make the wet-match-worlds not very dissimilar from the actual world, at all. But then what does explain why (22b), but not (2b), is felicitous? Here is an answer no one has given: the worlds where Sophie goes to the parade and is stuck behind someone tall are *just as similar* to the actual world as are the worlds where she goes to the parade and sees Pedro.[39] (Of course, it may be *unlikely* that Sophie would have been stuck behind someone tall, but

of closest worlds. In personal correspondence Karen Lewis has indicated that she rejects this possible revision [proposal] for several reasons.

[39] Karen Lewis also points out that "on a natural interpretation of a Lewis-style similarity ordering…the baseball player tripping and falling, and Sophie getting stuck behind someone tall at the parade count as among the closest worlds." (2016: 7-8) She uses this against the classic Lewisian semantics and in support of her contextualist account.

similarity is not a function of likelihood.[40]) In contrast, while the worlds at which the match was wet and struck could be very similar to the actual world indeed, they are *less* similar than the worlds at which the match was dry and struck (and lit).

If this is right, it means that (1a) <If Sophie had gone to the parade she would have seen Pedro> was never true (at least given the classic semantics). 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 the 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 that, for instance, Sophie might have been stuck behind someone tall, to make us aware of the possibility that she might not have seen Pedro.

There is good independent reason to think that the worlds at which Sophie went to the parade but was stuck behind someone tall are just as close as the worlds at which she went to the parade and saw Pedro. There are, I suggest, two tests that can be used to (defeasibly) determine if a particular set of antecedent-worlds are among the closest antecedent-worlds or not. Before introducing these tests it will be helpful to consider one last case.

[40] 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.

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

(24) a. It is a good thing 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, the child would not have broken a bone.
c. $_{ok}$Uhh, okay....but if that child had fallen, she would have broken a bone./$_{ok}$Okay, but if *that* child had fallen, she would have broken a bone./$_{ok}$Okay, but if that child, *just as she is*, had fallen, she would have broken a bone.

This is a very clear-cut case. The worlds where the child is wearing the padded body suit and falls are less similar to the actual world than worlds where she is wearing no padded suit and falls (note that this is the case whether 9 out of the 10 kids in the park are wearing padded suits, or whether there had earlier been a 0.9 chance that, that kid would be one of those chosen to wear the suit. All that matters is that in fact, that child is not wearing it). The various appropriate replies to (24b) (shown in (24c)) are suggestive of one kind of test, which can be dubbed the *Identify Something in the World Test*.

The "Identify Something in the World" Test: If some world, w1, is less similar to the actual world than is another world, w2, then there should be something one can 'point to' in the actual world in virtue of which this is so.

It is possible for the person who asserted (24a) to reply to (24b) 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 right now. The speaker can point to something about the world — the child, and what the child is wearing – to make it clear that the worlds where the child is wearing a protective suit are not relevant in the conversation. They are less similar worlds, and they are less similar in virtue of the fact that the child is actually dressed normally.

Sequence (23) passes the Identify Something in the World test as well.

(23) a. [Pointing to a match]: If this match had been struck it would have lit
b. If this match had been wet and struck it would not have lit.

In response to (23b) 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 to, 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 explain why worlds that are different in that respect are less close. Worlds where the match is not dry are less similar to the actual world in virtue of the fact that in the actual world the match is dry.

What about sequences (1) and (2)? What can be pointed to, in the actual world, which plausibly makes it such that the worlds where Sophie is stuck behind someone tall are less similar to the actual world than worlds where she sees Pedro? I say that there is nothing to point to. 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 not be infelicitous). 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 not be infelicitous). If there is nothing that can be pointed to, we conclude, by test one, that the worlds where Sophie attends the parade but is stuck behind someone tall are no further away than the worlds where she attends the parade and sees Pedro. Consider, now, a second test:

> **The "Probably" Test**: For two speakers S1 and S2, and any two counterfactuals (p1) A>C and (p2) A&B>not-C, if S1 asserts (p1) and, in reply, S2 asserts (p2), the nearest A&B worlds are no further from the actual world than the nearest A&not-B worlds only if it is appropriate for S1 to reply to S2's utterance with "you are right, what I should have said was 'A > probably-C'".[41]

Like the first test, I consider this test to be defeasible. I am not committed to it working in all cases (although I do not think it easy to generate counterexamples – I have not come up with any myself). But this test gives us another, basically reliable method for testing whether one world is as close to the actual world as another. The reason it works is that it is only appropriate to weaken one's assertion "A>C" to "A>probably-C" in response to the salience of "A&B>not-C" if A&B is a relevant, non-dismissible possibility. It is easiest to see this with some examples. We can again compare the original example with the match case and the falling child case:

(25) a. It is a good thing 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 wouldn't have broken a bone.
c. # You are right, what I should have said was 'if that child had fallen she probably would have broken a bone'.[42]

(26) a. [Pointing to a dry match]: If that match had been struck it would have lit
b. If that match had been wet and struck it would not have lit
c. #You are right, what I should have said was 'if that match had been struck it probably would have lit'.

[41] The test uses a conditional ("only if") rather than a biconditional ("if and only if") because there may be situations where it is unclear to those in the conversation whether a given world, or a set of worlds, is as similar to the actual world as another world, or another set of worlds. It is possible for the second sentence of a reverse Sobel sequence to sound infelicitous (and for the weakened "probably"-claim to sound felicitous) not because the possible worlds made salient by the first counterfactual are as close as the closest antecedent-worlds, but because the conversants do not *know* if they are or not. And if they do not know if they are or not, principles of epistemic humility can kick in to make infelicitous the second counterfactual in the reverse Sobel sequence.

[42] Although you might have the intuition that the speaker should have originally weakened his assertion with "probably" anyway (since, the speaker may not have actually been in position to say whether the child would have broken a bone or not), if it's the case that the speaker should have weakened his assertion in this way, it is certainly not because of what was said in (25b).

Now compare (27):

(27) a. If Sophie had gone to the parade she would have seen Pedro.
b. But if Sophie had gone to the parade and been stuck behind someone tall she would not have seen Pedro.
c. $_{\text{ok}}$You are right, what I should have said was 'if Sophie had gone to the parade she probably would have seen Pedro'.

(27c) is just the kind of thing that we'd expect someone to say in response to (27b). This indicates that worlds where Sophie is at the parade but stuck behind someone tall are relevant, and that they should be factored in when <if Sophie had gone to the parade she would have seen Pedro> is evaluated. In contrast, the infelicity of (26c) indicates that worlds where the match is wet and struck are not relevant to what is being expressed by (26a). Worlds where the match is wet are further away than worlds where it is dry, and so these worlds are dismissible. Their salience does *not* put pressure on the person asserting (26a) to weaken the assertion in order to take the previously ignored, wet-match-worlds into account.

I do not think it a coincidence that the sequences usually used to illustrate the problem of reverse Sobel sequences are sequences (1) and (2) (the counterfactuals in these sequences are not among those regularly used as examples in discussions of other counterfactual-related topics). This is not because I think that it has been widely recognized that worlds where Sophie is stuck behind someone tall are no further away than worlds where she sees Pedro. To the contrary, just about everyone in the literature has taken 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.[43] But it is the very fact

[43] Karen Lewis is, as far as I know, the only one who recognizes the possibility that the worlds where Sophie is stuck behind someone tall could be as similar to the actual world as worlds where Sophie sees Pedro (2017: xx). But she still assumes that (1a) is true: on her view worlds that are among the most similar can be 'pushed back' if not relevant, so

that, as it happens, *not* all nearest Sophie–goes–to–parade–worlds are Sophie–sees–Pedro–worlds that accounts for the infelicity of (2b).

This solution to the puzzle, according to which neither (1a) nor (2b) is true, avoids the problems faced by the alternative proposals. It explains why counterfactuals like (2b) seem fine when embedded in suppositional contexts or under attitude verbs: although (2b) is false (at least given the classic semantics), the speaker is supposing (or suspecting or fearing) that it is true.[44] And as we've seen, it can explain why felicitous sequences like (9) and (22) are felicitous, while (2) is not. Furthermore, it is compatible with would- and might-counterfactual duality: (1c) does not sound infelicitous, despite being incompatible with (1a) (given the duality thesis) because (1a) is untrue. If (1a) is untrue, and if (1b) brings this fact to salience, (1c) will sound just fine, despite being incompatible with (the untrue) (1a).

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 of the closest Sophie-goes-to-parade worlds are worlds where Sophie sees Pedro. This is not particular to this example. We've seen that in other cases (for example in the match case) when the worlds made salient by the first counterfactual in the reverse Sobel (i.e. A&not-C worlds) are not as similar as the most similar A&C worlds, the sequence is felicitous.[45]

(1a) can be true even if the worlds where Sophie is stuck behind someone tall are no less similar than worlds where Sophie sees Pedro.

[44] Or the speaker is being imprecise and suspecting or fearing that if Sophie had gone to the parade she would have *probably* seen Pedro. More on this shortly.

[45] This is not to say that the second counterfactual of a reverse Sobel sequence will *never* be infelicitous if it is true: I do not want to rule out the possibility that a counterfactual can sound infelicitous not because it is false but because the

If I am right, philosophers have been wrong to take reverse Sobel sequences as a challenge to the classic 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 serious difficulty for my proposal. Its truth entails that most ordinary counterfactuals are not true. Consider the match scenario, again. We know that worlds where the match is wet and struck are further away from the actual world than worlds where 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 where the match is struck but does not light. For instance, if someone had made salient the possibility that a gust of wind, or some indeterministic quantum event, could have occurred just as the match was being struck to make the match not light, this reverse Sobel sequence might be infelicitous:

(28) a. If the match had been struck but indeterministic quantum event *x* had occurred, it wouldn't have lit.
b. #But if that match had been struck it would have lit.

If it is a real (even if unlikely) possibility that quantum event *x* could have occurred and made the match not light, (28b) is infelicitous (in response to (28a), the person asserting (28b) should have at least weakened his assertion with something like 'probably'). And, as I have argued, if it is infelicitous, this is (generally[46]) because it is not true. It is not true because there are worlds

people in the conversation do not *know* if it is false. 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.

[46] See footnote 21.

just as close as the match-is struck-and-lights worlds where 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 counterfactuals, since as Hawthorne (2005), Hájek (manuscript), Lewis (2016) and others have pointed out, there is usually some world among the closest antecedent worlds where the consequent does not obtain for some reason or other.

How serious of a cost is this? Many would take it to be very serious. Many have gone to great lengths to avoid accepting the conclusion that most ordinary counterfactuals are not true.[47] I find the strong aversion to the claim somewhat surprising. We should not be so averse to thinking that most ordinary counterfactuals are not true for at least two reasons. The first is that people are almost always willing to modify their counterfactual assertion with "it is likely that" or "it is probable that", when pressed. In most cases it is not even necessary to bring a problematic possibility to salience to get a speaker to qualify his utterance in this way. A simple request for greater precision is usually enough. For instance, if someone says to me "if the match had been struck it would have lit", it will in most cases be enough for me to reply with "do you mean to say that if the match had been struck it would have *definitely* lit?" The natural response to this, I submit, is to retreat to something weaker: "Well, what I mean to say is that if the match had been struck it would have very likely lit", for instance. And if the speaker is not willing to retreat in this way, it is probably just for lack of imagination. If I confront her with reasons why it is possible that the match would not have lit (for instance, a gust have wind could have blown through the open

[47] See, for example, Lewis (1986), Bennett (2003), Williams (2008), Ichikawa (2011) and Lewis (2016), among many others.

window at just that moment!), it is only reasonable that she qualify her claim, then. (Of course, if there is no legitimate way for the match to have been struck but not lit – that is, if there are no antecedent worlds, among the closest, where the match does not light — then the unqualified counterfactual is true.)

Speakers' willingness to qualify their counterfactual assertions in this manner when pressed could suggest one of a couple of different things. It could mean that 'A>C' should be taken to mean something weaker than that all nearest A-worlds are C-worlds. Maybe 'A>C' means that *most* nearest A–worlds are C–worlds, or that almost all nearest A–worlds are C–worlds. The second possibility is that 'A>C' *does* mean that all nearest A–worlds are C–worlds, but that when we use the counterfactual (unmodified) we are simply being imprecise: if speaking precisely we'd say that 'A>probably–C'.[48] There are several reasons to go with the second option over the first. One is that one's credence in a given counterfactual A>C is generally not the same as one's credence in its weaker counterpart. I might have credence of 0.8 that if Sophie had gone to the parade she would have seen Pedro, but credence very near 1.0 that if Sophie had gone to the parade she would have probably seen Pedro. This suggests that the two do not mean the same thing.[49] Another reason to deny that 'A>C' means that almost all A–worlds are C–worlds was already given in footnote 16: treating 'A>C' in this way commits us to accepting that any event sufficiently likely to have happened had the antecedent obtained, *would* have happened, had the antecedent obtained (where

[48] Hájek endorses something like this in his (ms).

[49] Note that if David Lewis's semantics is right, my credence in <If Sophie had gone to the parade she would have seen Pedro> should actually probably be much lower than 0.8, since it seems quite likely that there is at least one nearest Sophie-goes-to-the-parade-world where Sophie does not see Pedro (in which case, on David Lewis's semantics, the counterfactual is false). Stalnaker (1981) and Edgington (1995), (2014) have objected to the Lewisian semantics on these sorts of grounds.

by “sufficiently likely” I mean above the threshold to make the counterfactual true on this view, whatever that threshold is). But since the extremely likely does not always happen in the actual world, there is little justification for thinking that the extremely likely would have always happened counterfactually.

A much better explanation for why speakers are usually willing to retreat to a more modest “A>*probably*–C” when pressed is that when in a thoughtful mood, speakers can come to realize that “A>C” is just too strong. If the speaker wants to communicate what she (can, upon reflection, come to realize she) means more precisely, she must say something weaker. The fact that speakers are usually happy to admit that they meant something a bit weaker than what is captured by ‘*all* nearest A–worlds are C–worlds’, is one reason not to be reluctant to accept the possibility that most unqualified counterfactuals are not true. A second reason not to be averse to the possibility is that imprecision is utterly ubiquitous in ordinary language.[50] When I say that I am 5’6” tall I do not mean that I am precisely 5’6’’ tall. I am speaking loosely. What I really mean, which I will readily admit to if pressed, is that my height is somewhere in the vicinity of 5’6’’. I am *about* 5’6’’ tall. To take another ordinary example, suppose that someone is baking some 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 likely smells subtly different than what Jasper perceives. What Jasper smells is actually a mixture of bread and a host of other things combined. And what of the chemist who says that water is H_2O? Does she say something true? Maybe not, technically. Here is Paul Teller (forthcoming):

[50] On this see Teller (2011), Braun and Sider (2007), and Elgin (2004).

> Water is never absolutely 100% a collection of H_20 molecules. There are always some impurities and some dissociated molecules. When one closely approaches the (in practice unobtainable) goal of 100% H_20 the properties of the substance become significantly different from the water of familiar experience. So, more carefully, water is MOSTLY H_20. (Teller's emphasis)

The examples are endless. It is not so radical to hold that people regularly speak loosely when asserting counterfactuals if people regularly speak loosely in general.

In addition to not always quite speaking the truth we often speak more confidently than what might reasonably be called for. A large proportion of what we say can, and perhaps should, be appropriately weakened with "probably": "He's at the park" can be "he's probably at the park" (he said he was going to the park, but he could have lied or changed his mind); "the car is in the lot" can be "the car is probably in the lot" (it could have been towed or stolen), etc. Frequently we leave the 'probably' out although we'd usually be happy to put it in if pressed. It is no surprise, then, that we do the same when speaking counterfactually.[51]

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 the smell he smells is the smell of bread. For one, Jasper might not in that moment recognize that there are likely to be other smells adulterating the otherwise pure bread smell. And even if he were to realize that bread probably does not smell *exactly* like *that*, he would risk misleading his listeners if he were to say something

[51] Of course the difference between non-counterfactual unqualified assertions and counterfactual unqualified assertions is that the former are still *true* if the speaker is correct (even if he spoke too confidently) whereas if I am right, counterfactual utterances are usually *not* true if not appropriately qualified. But this is just a consequence of the kind of strange creatures counterfactuals are. In the case of non-counterfactual propositions like "he is in the park", the world (or something in it) makes the proposition true if he is actually in the park, even if the speaker was not in an epistemic position to appropriately assert the sentence. In contrast, there is nothing in the actual world to make the unqualified counterfactual true if, given the way the world is, there are multiple, equally good candidates for how it could have been. But given that ordinary speakers are not in general privy to these differences, we should not expect them to be more concerned about not qualifying their counterfactual assertions than they are about not qualifying their non-counterfactual assertions.

more precise. If, just to be safe, Jasper were to say "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 than he does. For instance, he might implicate that he recognizes 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 closer to the truth in the first place – e.g., 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, and *not* unreasonably, that I had some particular reason to think that the match may not have lit if struck. Maybe I know something about the match, or its environment, that they do not know. 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 the odds to be. Since speakers generally do not weaken their counterfactual – or, for that matter, their non-counterfactual – assertions in this way, to do so could suggest some degree of uncertainty that may not in fact be felt. For these reasons, and because it can be tiresome and unnecessary for speakers to be precise, and because the 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'. And if it is something slightly different than A>C that is (upon reflection) intended, it should not so much matter if A>C is not true.

VI. Conclusion

I have argued that we should not be averse to accepting that most unqualified counterfactuals are not true. They are close to something that is true, and that is good enough. As for the reverse Sobel sequence "problem", it is not really a problem at all. Once we are willing to accept that most unqualified counterfactuals are not true, we can recognize infelicitous reverse Sobel sequences for what they are: evidence that we've come across a counterfactual that should be qualified. If this is right, there is no need to revise or reject the classic semantics in response to reverse Sobel sequences. Nevertheless, it might prove worthwhile to shift some attention away from trying to understand unqualified counterfactuals first and foremost, and toward trying to advance our understanding of the qualified ones.

REFERENCES

Ahmed, A. (2011). "Out of the Closet", *Analysis* 77-85.

Barker, S. (1999), "Counterfactuals, Probabilistic Counterfactuals, and Causation," *Mind* 108: 427-469.

Bennett, J. (2003), *A Philosophical Guide to Counterfactuals*, Oxford University Press.

Braun, D. & Sider, T. (2007) "Vague, So Untrue," *Nous* 41 (2): 133-156.

Edgington, D. (1995) "On Conditionals," *Mind*, 104: 235-329.

Edgington, D. (2004). "Counterfactuals and the Benefit of Hindsight". In *Causation and Counterfactuals*, eds P. Dowe and P. Noordhof, 12–27. London: Routledge.

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

Edgington, D. (2014) "Indicative Conditionals", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2014/entries/conditionals/>.

Elgin, C. (2004) "True Enough", *Philosophical Issues*, 14: 113-121.

Gillies, T. (2007) "Counterfactual Scorekeeping," *Linguistics and Philosophy*, vol. 30: 329-360.

Goodman, N. (1946). "The Problem of Counterfactual Conditionals", Reprinted in Goodman, *Fact, Fiction, and Forecast*, 4th ed. Cambridge: Harvard, 1979.

Hájek, A. (manuscript), *Most Counterfactuals are False*. ANU, monograph in progress.

Hawthorne, J. (2004) *Knowledge and Lotteries*, Oxford: Clarendon Press.

Hawthorne, J. (2005) "Chance and Counterfactuals", *Philosophy and Phenomenological Research*, Vol. LXX, No. 2.

Hiddleston, E. (2005). "A Causal Theory of Counterfactuals", Nous 39:4.

Ichikawa, J. (2011) "Quantifiers, Knowledge, and Counterfactuals," *Philosophy and Phenomenological Research*, Vol. LXXXII No. 2.

Kment, B. (2010) "Causation: Determination and Difference-Making", *Nous* 44:1, pp. 80-111.

Leitgeb, H. (2012a) A Probabilistic Semantics for Counterfactuals: Part A," *The Review of Symbolic Logic* 5: 1, 16-84.

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

Lewis, D. (1973) *Counterfactuals*. Basil Blackwell Ltd., Malden, MA.

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

Lewis, D. (1986) "Causation", reprinted in *Philosophical Papers, Vol. II*, New York: Oxford University Press, pp. 159-213.

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

Lewis, K. (work in progress) "Counterfactual Discourse in Context", posted at http://www.columbia.edu/~kl2663/.

Moss, S. (2012), "On the Pragmatics of Counterfactuals", *Nous* 46:3, 561-586.

Noordhof, P. (2005). "Morgenbesser's Coin, Counterfactuals and Independence", *Analysis* 65: 261-63.

Phillips, I. (2007). "Morgenbesser Cases and Closet Determinism". Analysis 67: 42–49.

Pruss, A. (2003). "David Lewis's Counterfactual Arrow of Time", *Nous* 37:4, pp. 606-637.

Sartorio, C. (2005) "Causes as Difference-Makers", *Philosophical Studies* 123: 71-96.

Schaffer, J. (2000) "Trumping Preemption", *The Journal of Philosophy* 97:4, pp. 165-181.

Schaffer, J. (2004). "Counterfactuals, Causal Independence and Conceptual Circularity", *Analysis* 64: 299–309.

Schaffer, J. (2005). "Contrastive Causation", *The Philosophical Review*, Vol. 114, No. 3.

Stalnaker, R. (1968). "A Theory of Conditionals," In Harper et al. (1981), 41-55.

Stalnaker, R. (1981). "A Defense of Conditional Excluded Middle," in Harper, Stalnaker and Pearce (eds.), pp. 87-104.

Teller, P. (2011), “Two Models of Truth,”, *Analysis* 71:3, 465-472.

Teller, P. (forthcoming), “Truth and Fiction in Science,” *Science et Avenir.*

Von Fintel, K. (2001) “Counterfactuals in a Dynamic Context,” in *Ken Hale: A Life in Language*, Michael Kenstowicz, editor, MIT Press, Cambridge.

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

Won, C. (2009). “Morgenbesser’s Coin, Counterfactuals, and Causal Versus Probabilistic Independence”. *Erkenntnis* 71: 345-354.