Noûs

NOÛS 46:3 (2012) 561–586

On the Pragmatics of Counterfactuals

Sarah Moss
University of Michigan

Stalnaker 1968 and Lewis 1973 advocate a certain semantics for counterfactuals, conditionals such as:

(1) If Sophie had gone to the New York Mets parade, she would have seen Pedro Martínez.

Until recently, theirs was the standard theory. But von Fintel 2001 and Gillies 2007 present a problem for the standard semantics: they claim that it fails to explain the infelicity of certain sequences of counterfactuals, namely *reverse Sobel sequences*. Both von Fintel and Gillies propose alternative dynamic semantic theories that explain the infelicity of reverse Sobel sequences, and argue that we should trade in the standard semantics of counterfactuals for theirs.

I will argue that we can explain the infelicity of reverse Sobel sequences without giving up the standard semantics. In §1, I present the Stalnaker-Lewis semantics. In §2, I introduce reverse Sobel sequences, discuss the von Fintel and Gillies theories, and say how their theories predict the infelicity of reverse Sobel sequences. In §3, I give my own explanation of why reverse Sobel sequences are generally infelicitous. In §4, I argue that my independently motivated pragmatic theory accounts for a range of subtle judgments about sequences of counterfactuals. For instance, I argue that some reverse Sobel sequences are *felicitous*, and that my theory gives a successful account of our judgments about these sequences. Finally, in §5, I discuss another potential application of my approach: infelicitous sequences containing ‘might’ counterfactuals.

Developing pragmatic theories of counterfactuals is not just of intrinsic interest. Reverse Sobel sequences have been cited as evidence in favor of the *dynamic turn*: a paradigm shift away from semantic theories that assign truth conditions to utterances, to theories that assign rules for updating contexts.[1] Hence getting clear about the pragmatics of these sequences should help us make progress in a key debate about the fundamental units of linguistic meaning.

1 Sobel sequences and the standard semantics

Before too much thinking, it is tempting to say that (1) is true just in case all possible worlds in which Sophie goes to the parade are worlds in which she sees Pedro. This is the *strict conditional analysis* of counterfactuals. On this analysis, the context in which a counterfactual is uttered contributes a function to its truth conditions: an accessibility function *f* from worlds to sets of worlds. 'If *p*, would *q*' then expresses a proposition that is true at a world just in case all the *p* worlds that are *f*-accessible from that world are *q* worlds.[2]

But now consider the following sequence of counterfactuals:

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

Intuition says that the counterfactuals in (2) can be true together. But the strict conditional analysis predicts otherwise. For on this analysis, (2a) says that all possible worlds in which Sophie goes to the parade are worlds in which she sees Pedro. Given that there are possible worlds in which Sophie goes to the parade and is stuck behind a tall person, this is incompatible with what (2b) says, namely that all possible worlds in which Sophie goes to the parade and is stuck behind a tall person are worlds in which she does not see Pedro. So the strict conditional analysis predicts that (2a) and (2b) cannot be true together.

Sequences like (2) are *Sobel sequences*.[3] LEWIS 1973 made Sobel sequences famous, and was motivated by them to reject the strict conditional analysis of counterfactuals. On both the Lewis analysis and its cousin in STALNAKER 1968, the context in which a counterfactual is uttered contributes a similarity ordering *O* on worlds to its truth conditions, rather than contributing a function on worlds. Roughly speaking, 'if *p*, would *q*' expresses a proposition that is true at a world just in case all the *p* worlds closest-by-*O* to that world are *q* worlds.[4] Stalnaker and Lewis predict that the counterfactuals in (2) can both be true. For according to them, (2a) says that the closest worlds in which Sophie goes to the parade are worlds in which she sees Pedro, and that is perfectly compatible with what (2b) says, namely that the closest worlds in which Sophie goes to the parade and is stuck behind a tall person are worlds in which she does not see Pedro. Hence this analysis looks promising, and until recently, most theorists accepted some version of this analysis of counterfactuals.

2 Reverse Sobel sequences

But VON FINTEL 2001 and GILLIES 2007 raise a problem for the standard analysis of counterfactuals.[5] Suppose we reverse the order of the sentences in (2) to make the following sequence (3):

(3a) If Sophie had gone to the parade and been stuck behind a tall person, she would not have seen Pedro.
(3b) #But if she had gone to the parade, she would have seen Pedro.

Both von Fintel and Gillies say that when uttered in this order, if (3a) is true then (3b) is not true. But according to von Fintel and Gillies, the standard analysis predicts otherwise. For on the standard analysis, the order in which the counterfactuals in (2) and (3) are uttered makes no difference to their semantic value. Even if you have just said that the closest worlds in which Sophie goes to the parade and is stuck behind a tall person are ones where she does not see Pedro, you may go on to truly say that the closest worlds in which she goes to the parade are ones where she sees Pedro, with no fear of contradiction. So the standard analysis predicts that even when uttered in sequence (3), the counterfactual (3b) can be true.

Sequences like (3) are *reverse Sobel sequences*. These sequences motivate von Fintel and Gillies to trade in the standard analysis of counterfactuals for another theory. Surprisingly, they trade in the standard analysis for a variant on the original strict conditional analysis of counterfactuals.[6] In other words, they want to preserve the original claim that 'if *p*, would *q*' is true just in case all the *p* worlds in a contextually determined set are *q* worlds. But they augment this claim with a strong claim about the dynamics of conversation: as part of their meaning, counterfactuals effect changes in the context. In particular, counterfactuals impose demands on the contextually determined domain that subsequent counterfactuals quantify over.[7]

In (2001), von Fintel endorses much of the strict conditional analysis. He adopts the claim that context contributes an accessibility function *f* to the truth conditions of counterfactuals, a function from worlds to sets of worlds. He adopts the claim that 'if *p*, would *q*' expresses a proposition that is true at a world just in case all the *p* worlds that are *f*-accessible from that world are *q* worlds. But von Fintel adds that there is a second contextual parameter relevant to the interpretation of counterfactuals: a similarity ordering on worlds. He also adds that there is another component to the meaning of a counterfactual: its effect on the accessibility function *f*. In particular, 'if *p*, would *q*' demands that from every world, there be some *f*-accessible *p* worlds.

More precisely, von Fintel says that counterfactuals "update *f* by adding to it for any world *w* the closest antecedent worlds" (20). Suppose that the accessibility function of some context maps some world *w* to a set that contains no p worlds. Uttering ‘if p, would q’ in that context updates the accessibility function, so that it maps that same world w to a set that does contain p worlds. In particular, the updated function maps w to the set of all the worlds at least as close to w as the nearest p worlds, by the contextually determined similarity ordering.

I have said that according to von Fintel, ‘if p, would q’ *demands* that from every world, there be some f-accessible p worlds. There are several ways to understand the nature of this demand. On one version of the proposal, the semantic value of ‘if p, would q’ is a pair of update rules: a rule for updating the context set and a rule for updating the accessibility function f. On another version, it is part of the meaning of ‘if p, would q’ that it presupposes that from every world, there are f-accessible p worlds. On either version, the upshot is the same: once ‘if p, would q’ is asserted, there must be p worlds in the domain that the counterfactual quantifies over.[8]

This dynamic analysis predicts that (3b) must be false, while (2b) can be true. (3a) demands that there be some accessible worlds in which Sophie goes to the parade and is stuck behind a tall person. So once we meet the demands of (3a), there are some accessible worlds in which Sophie goes to the parade and does not see Pedro. (3b) says that Sophie sees Pedro in all accessible worlds in which she goes to the parade. So once we utter (3a) and accommodate its demands, (3b) must be false. But Sobel sequences do not crash in the way that reverse Sobel sequences do. (2a) demands merely that there are some accessible worlds in which Sophie goes to the parade. (2b) says that in all accessible worlds in which Sophie goes to the parade and is stuck behind a tall person, she does not see Pedro. So even once we utter (2a) and accommodate its demands, (2b) can be true.

One might prefer a slight variation on this analysis. In his (1997), von Fintel argues that ‘if p, would q’ presupposes a local application of the law of conditional excluded middle: that either all or none of the accessible p worlds are q worlds. If you accept this claim, and also accept that a counterfactual lacks a truth value when this particular presupposition is false, then given the dynamic semantics presented above, you will conclude that (3b) is not false, but merely lacks a truth value. My arguments concerning the dynamic approach apply equally to this analysis of reverse Sobel sequences.

Gillies 2007 develops a dynamic semantics of counterfactuals similar to von Fintel’s. On von Fintel’s analysis, there are two contextual parameters: a regularly updated accessibility function, and a similarity ordering on worlds. Gillies posits only one parameter: a *counterfactual hyperdomain*, i.e. a collection of nested sets of worlds. He says that ‘if p, would q’ is true just in case all the p worlds in the smallest set in the counterfactual hyperdomain are q worlds. Gillies then adds another component to the meaning of a counterfactual: ‘if p, would q’ demands that there are some p worlds in the smallest set in the counterfactual hyperdomain.

Gillies predicts that (3b) cannot be true if (3a) is true, in almost exactly the same way von Fintel does. Once we accommodate the demands of (3a), there are some worlds in the smallest set in the counterfactual hyperdomain in which Sophie goes to the parade and does not see Pedro. (3b) says that Sophie sees Pedro in all the worlds in the smallest set in the counterfactual hyperdomain. So once we utter (3a) and accommodate its demands, (3b) cannot be true. GILLIES 2007 concludes that reverse Sobel sequences are *inconsistent*: a reverse Sobel sequence "cannot be interpreted without collapse into absurdity" (28). But the demands of (2a) are weaker, so even once we accommodate them, (2b) can be true.

Besides VON FINTEL 2001 and GILLIES 2007, I know of only one other analysis of conditionals that aims to account for phenomena like the infelicity of (3). WILLIAMS 2008 observes that the indicative analog of (2) is felicitous:

(2a′) If Sophie went to the parade, she saw Pedro.
(2b′) But if Sophie went to the parade and got stuck behind a tall person, she did not see Pedro.

Meanwhile, the indicative analog of (3) is infelicitous:

(3a′) If Sophie went to the parade and got stuck behind a tall person, she did not see Pedro.
(3b′) #But if Sophie went to the parade, she saw Pedro.

Williams accounts for these data by adopting a variant of the strict conditional analysis for indicative conditionals, according to which the domain of the necessity modal is the *context set*: the set of worlds compatible with what is treated as true for purposes of conversation. He says that 'if *p*, *q*' is true just in case all *p* worlds in the context set are *q* worlds. Like von Fintel and Gillies, Williams then adds another component to the meaning of a conditional: Williams says that 'if *p*, *q*' presupposes that the context set contains some *p* worlds.

It is not clear how to generalize Williams' theory to an analysis of counterfactuals. It is okay to utter (1) even if you know that Sophie did not go to the parade. In general, it is okay to utter a counterfactual even if the antecedent is presupposed to be false. So the *counterfactual* conditional 'if *p*, would *q*' does not presuppose that the context set contains some *p* worlds. Williams does not spell out an analysis of counterfactuals. But he says that the analysis of counterfactuals in GILLIES 2007, though developed independently, is "similar in spirit" to his own theory. Insofar as the generalization of Williams' theory to counterfactuals resembles the analysis in GILLIES 2007, my arguments concerning Gillies apply to Williams too.

To sum up how things stand so far: the strict conditional analysis of counterfactuals says that 'if *p*, would *q*' is true just in case all the possible *p* worlds are q worlds. Sobel sequences motivate Lewis to reject this story for an analysis according to which 'if p, would q' is true just in case all the closest possible p worlds are q worlds. Reverse Sobel sequences motivate von Fintel and Gillies to reject this story for a variant of the strict conditional analysis, according to which 'if p, would q' is true just in case all the p worlds in a certain contextually determined domain are q worlds, and the same sentence demands that there be some p worlds in that domain.

Presented with these theories, one might be tempted to revive the standard analysis. Strictly speaking, the standard analysis can accommodate the infelicity of reverse Sobel sequences. The second sentence of a reverse Sobel sequence might be false because the similarity ordering determined by context changes when you utter the first sentence. The advocate of the standard analysis can say that uttering (3a) changes the contextually determined similarity ordering, expanding the set of closest worlds in which Sophie goes to the parade until it includes some worlds in which Sophie is stuck behind a tall person. (3b) would be false as uttered after such a context change.

Of course, it is not exactly in the spirit of the standard analysis to think that it is so easy to change the contextually determined similarity ordering. Lewis and Stalnaker explain why one can truly utter (2b) after (2a) without saying that the contextually determined similarity ordering changes in (2). They simply say that according to the single similarity ordering in play throughout (2), worlds where Sophie is stuck behind a tall person are farther away than some other worlds where she goes to the parade. Given that Lewis and Stalnaker do not posit changes in the similarity ordering to explain (2), it seems against the spirit of the standard analysis to posit such changes to explain (3).

But at this point in the game, von Fintel and Gillies can claim a greater advantage over the standard semantics: the dynamic approach is a stronger theory, yielding systematic predictions about when counterfactuals are felicitous. For example, it is part of the dynamic semantic value of (3a) that it effects particular changes on the domain that counterfactuals quantify over. So the dynamic theory itself entails that (3b) will be infelicitous after (3a) is uttered. Lewis and Stalnaker may say that (3b) is infelicitous in contexts such that Sophie gets stuck behind a tall person in some of the closest possible worlds in which she goes to the parade. But nothing in their theory predicts that uttering (3a) will make the context be this way. Lewis 1973 in fact dismisses a version of the *strict conditional analysis* on similar grounds. Lewis considers only judgments about Sobel sequences, not reverse Sobel sequences. He says that appealing to context shifting in order to explain the felicity of Sobel sequences is "defeatist…consign[ing] 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" (13).

Gillies 2007 responds to Lewis:

> To see that this kind of story is not the stuff of defeatism we only have to see that the interaction between context and semantic value, mediated by a mechanism of local accommodation, can be the stuff of formal and systematic analysis. To see that this is not a mere loophole, we only have to see that facts about counterfactuals in context—the discourse dynamics surrounding them—are best got at by the kind of story I want to tell. (2)

To sum up: Gillies and von Fintel claim that positing changes in the similarity ordering is the only way for the standard semantics to account for the infelicity of reverse Sobel sequences. If that were true, then on behalf of advocates of the standard semantics, I would concede: we should prefer a theory that yields systematic predictions about when counterfactuals are felicitous. The point of the dynamic approach is to provide just this kind of theory.

3 A pragmatic account of reverse Sobel sequences

However, I think von Fintel and Gillies are wrong: the standard semantics can account for the infelicity of reverse Sobel sequences, without positing changes in the similarity ordering. In this section, I will give an alternative explanation for the infelicity of reverse Sobel sequences. My explanation is compatible with a Stalnaker-Lewis analysis on which uttering sequences like (2) and (3) does not change the contextually determined similarity ordering.

Suppose we are enjoying a perfectly normal day at the zoo, looking at an animal in the zebra cage that seems to have natural black and white stripes. It has not recently crossed our minds that the zoo may be running a really low-budget operation, where they paint mules to look like zebras. In this situation, I might have reason to say:

(4a) That animal was born with stripes.

If you are in a slightly pedantic mood, you might reply with the following:

(4b) But cleverly disguised mules are not born with stripes.

This reply may be a non sequitur, perhaps even a little annoying. But otherwise, there is nothing wrong with your reply. On the other hand, once you have mentioned cleverly disguised mules, I would not be willing to repeat my original assertion. I may even feel as if I ought to take back what I said. In other words, there is a contrast between sequence (4) and the following sequence:

(5a) Cleverly disguised mules are not born with stripes.
(5b) #But that animal was born with stripes.

I would like to suggest that Sobel sequences are okay for the same reason (4) is, and reverse Sobel sequences are bad for the same reason (5) is.[9]

So why is (5) bad, while (4) is okay? Here is one intuitive answer: in the above scenario, (5b) is infelicitous because (5a) raises the possibility that the caged animal is a cleverly disguised mule, and the speaker of (5b) cannot rule out this possibility. So (5b) is infelicitous because in the above scenario, it is an epistemically irresponsible thing to say. Meanwhile, it is perfectly okay to utter the same sentences in the reverse order, since uttering (4a) is not epistemically irresponsible when no recherché possibilities are salient, and (4a) does not raise possibilities to salience that the speaker of (4b) irresponsibly ignores.

Our intuitions about (5) point towards a general principle governing assertability.[10]

(EI) It is epistemically irresponsible to utter sentence S in context C if there is some proposition ϕ and possibility μ such that when the speaker utters S:

(i) S expresses ϕ in C
(ii) ϕ is incompatible with μ
(iii) μ is a salient possibility
(iv) the speaker of S cannot rule out μ.

(EI) tells us that if a speaker cannot rule out a possibility made salient by some utterance, then it is irresponsible of her to assert a proposition incompatible with this possibility.[11] Hence we can use (EI) to explain why it is infelicitous to utter (5b) in the scenario described above. It simply remains to be shown that we can use this independently motivated principle to explain why it is generally infelicitous to utter reverse Sobel sequences.

Earlier I stipulated that the speaker of (5b) could not rule out that a certain animal was a cleverly disguised mule. One can make a similar claim about reverse Sobel sequence scenarios: the speaker of the second sentence of a reverse Sobel sequence generally cannot rule out certain possibilities incompatible with the content of her utterance. Given (EI), this explains why it is generally infelicitous to utter the second sentence of a reverse Sobel sequence.

For example, consider again the reverse Sobel sequence:

(3a) If Sophie had gone to the parade and been stuck behind a tall person, she would not have seen Pedro.
(3b) #But if Sophie had gone to the parade, she would have seen Pedro.

Someone who utters (3b) generally will not be able to rule out the possibility that if Sophie had gone to the parade, she might have been stuck behind a tall person. Hence (EI) entails that it is epistemically irresponsible to utter (3b), since:

(i) (3b) expresses the proposition that Sophie would have seen Pedro if she had gone to the parade.
(ii) The proposition that Sophie would have seen Pedro if she had gone to the parade is incompatible with the possibility that Sophie might have been stuck behind a tall person if she had gone to the parade.
(iii) The possibility that Sophie might have been stuck behind a tall person if she had gone to the parade is a salient possibility.
(iv) The speaker of (3b), at the time at which she utters (3b), cannot rule out the possibility that Sophie might have been stuck behind a tall person if she had gone to the parade.

The same goes for other reverse Sobel sequences: ‘if *p* and *r*, would not-*q*’ raises a certain possibility to salience, namely that *r* might have been the case if *p* had been the case. Ordinarily it simply does not take much to raise this possibility to salience; often merely mentioning the possibility that *r* suffices. Furthermore, the speaker who then utters ‘if *p*, would *q*’ generally cannot rule out this possibility. Finally, the speaker of ‘if *p*, would *q*’ expresses a proposition incompatible with this possibility. For this reason, it is generally infelicitous to utter the second sentence of a reverse Sobel sequence: it is epistemically irresponsible to assert a proposition incompatible with an uneliminated possibility that the first sentence raises to salience.[12]

This way of applying (EI) to a reverse Sobel sequence case depends on the following: that ‘if *p*, would *q*’ expresses a proposition incompatible with the possibility that *r* might have been the case if *p* had been the case. To make this more precise: even once the second sentence of a reverse Sobel sequence is uttered, it is still an accepted background fact that if *p* and *r*, would not-*q*. In Lewis’s logic of counterfactuals, we can derive a contradiction from this proposition, together with the proposition expressed by ‘if *p*, would *q*’ and the possibility that if *p*, might *r*.

Here are the relevant rules and axioms of **VC**, Lewis’s official logic of counterfactuals:[13]

rule 2: *Deduction within conditionals*: for any $n \geq 1$,

$$\frac{\vdash (\chi_1 \wedge \cdots \wedge \chi_n) \supset \psi}{\vdash ((\phi \mathrel{\Box\!\!\rightarrow} \chi_1) \wedge \cdots \wedge (\phi \mathrel{\Box\!\!\rightarrow} \chi_n)) \supset (\phi \mathrel{\Box\!\!\rightarrow} \psi)}$$

axiom 1: *Truth-functional tautologies*
axiom 2: *Definitions of non-primitive operators*
axiom 5: $(\phi \mathrel{\Box\!\!\rightarrow} \neg\psi) \vee (((\phi \wedge \psi) \mathrel{\Box\!\!\rightarrow} \chi) \equiv (\phi \mathrel{\Box\!\!\rightarrow} (\psi \supset \chi)))$

Using these rules and axioms, we can derive a contradiction from the proposition expressed by ‘if *p*, would *q*’ and the salient possibility that if *p* might *r*, as follows:

1.	$p \Diamond\!\!\rightarrow r$	salient possibility
2.	$\neg(p \Box\!\!\rightarrow \neg r)$	1, axiom 2
3.	$(p \wedge r) \Box\!\!\rightarrow \neg q$	background facts
4.	$(p \Box\!\!\rightarrow \neg r) \vee (((p \wedge r) \Box\!\!\rightarrow \neg q) \equiv (p \Box\!\!\rightarrow (r \supset \neg q)))$	axiom 5
5.	$p \Box\!\!\rightarrow (r \supset \neg q)$	2, 3, 4, axiom 1
6.	$((r \supset \neg q) \wedge q) \supset \neg r$	axiom 1
7.	$((p \Box\!\!\rightarrow (r \supset \neg q)) \wedge (p \Box\!\!\rightarrow q)) \supset (p \Box\!\!\rightarrow \neg r)$	6, rule 2
8.	$\neg(p \Box\!\!\rightarrow q)$	2, 5, 7, axiom 1
9.	$p \Box\!\!\rightarrow q$	expressed proposition
10.	$\bot$	8, 9, axiom 1

The second sentence of a reverse Sobel sequence expresses the proposition that if *p*, would *q*. For example, (3b) expresses the proposition that if Sophie had gone to the parade, she would have seen Pedro. In order to explain how this leads to a violation of the (EI) conditions, one must examine what is *common ground* when (3b) is uttered, in the sense of STALNAKER 2002: the information that conversational participants take for granted as background information for purposes of conversation. The notion of incompatibility relevant to condition (ii) of (EI) is incompatibility of the relevant propositions, given what is common ground when a speaker asserts the content in question.[14] Since the proposition expressed by (3a) is common ground when (3b) is uttered, one can use what is common ground when (3b) is uttered to derive a contradiction from the proposition expressed by (3b) and the possibility that was raised by (3a), as outlined above. Since the asserted content of (3b) can thereby be shown to be incompatible with a live salient possibility, the conditions of (EI) are violated and (3b) is infelicitous.

One cannot similarly use what is common ground when (2b) is uttered to derive a contradiction from the proposition expressed by (2b) and the possibility raised by (2b) itself. (2a) expresses the proposition that if Sophie had gone to the parade, she would have seen Pedro. But this proposition is intuitively no longer common ground once (2b) is uttered. In the zebra examples, there is a similar asymmetry in whether the proposition expressed by the first sentence typically remains common ground throughout the sequence. Consider the zebra sequences again:

(4a) That animal was born with stripes.
(4b) But cleverly disguised mules are not born with stripes.

(5a) Cleverly disguised mules are not born with stripes.
(5b) #But that animal was born with stripes.

Once the speaker of (4b) says that cleverly disguised mules are not born with stripes, it is typically no longer common ground that the caged animal under discussion was born with stripes. So one cannot use what is common ground when (4b) is uttered to derive a contradiction from the proposition expressed by (4b) and the possibility that the animal in question is a cleverly disguised mule. But even once the speaker of (5b) says that the caged animal was born with stripes, it is typically still common ground that cleverly disguised mules are not born with stripes.

Our natural responses to these examples are independent evidence of this asymmetry in the resilience of common ground information. For instance, it is natural to respond to (5b) by saying:

(5c) But how do you know that this animal was born with stripes? After all, we said that mules are not born with stripes, and for all you know, this animal might be a mule.

But an analogous response to (4b) typically sounds bad:

(4c) #But how do you know that mules are *not* born with stripes? After all, we said that this animal was born with stripes, and for all you know, this animal might be a mule.

And the analogous response to (2b) sounds similarly bad:

(2c) #But how do you know that if Sophie had gone and been stuck behind a tall person, she would have missed Pedro? After all, we said that if she had gone, she would have seen Pedro, and for all you know, if she had gone, she might have been stuck behind a tall person.

I will not defend any general theory of how the common ground of a conversation behaves under various conversational pressures. Ultimately, what matters for my purposes is not the exact nature of the mechanism at work in these examples. I am simply interested in general arguments concerning whether this mechanism is semantic or pragmatic in nature.

So far I have spelled out one way to derive a contradiction from a salient possibility and the proposition expressed by the second sentence of a reverse Sobel sequence. There are other ways to apply (EI) to a reverse Sobel sequence case. The second step of the above derivation appeals to axiom 2 of **VC**, and in particular to Lewis's definition of the 'might' operator. For Lewis, 'might' and 'would' counterfactuals are duals:

$$(\phi \diamond\!\!\rightarrow \psi) \equiv \neg(\phi \,\Box\!\!\rightarrow \neg\psi)$$

However, this duality thesis is a contentious assumption. For now, I wish to remain neutral about the duality of 'might' and 'would' counterfactuals. If you reject the duality thesis, there are other ways to derive a contradiction from a salient possibility and the proposition expressed by ‘if *p*, would *q*’. For instance, you might think that the first sentence of a reverse Sobel sequence raises the possibility that it is not the case that if *p* were the case, then not-*r* would be the case, even though you reject that this possibility is equivalent to the possibility that if *p* were the case, *r* might be the case.

Alternatively, you may be one of many theorists who are motivated to reject the duality thesis in order to accept the law of conditional excluded middle.[15] In other words, you may think that one of the following must hold: that if *p* were the case, then *r* would be the case, or that if *p* were the case, then not-*r* would be the case. In that case, you will likely think that it does not take much to raise the possibility that the first of these is true, and you will likely accept that ‘if *p* and *r* were the case, then not-*q* would be the case’ raises the possibility that if *p* were the case, then *r* would be the case. Given that there are possible *p* worlds, it is again possible to derive a contradiction between this salient possibility and the proposition expressed by the second sentence in the reverse Sobel sequence. Simply replace steps 1–2 of the above derivation with the following:

1a.	$p \mathrel{\Box\!\!\rightarrow} r$	salient possibility
1b.	$(\neg r \wedge r) \supset \bot$	axiom 1
1c.	$((p \mathrel{\Box\!\!\rightarrow} \neg r) \wedge (p \mathrel{\Box\!\!\rightarrow} r)) \supset (p \mathrel{\Box\!\!\rightarrow} \bot)$	1b, rule 2
1d.	$\neg(p \mathrel{\Box\!\!\rightarrow} \bot)$	assumption
2.	$\neg(p \mathrel{\Box\!\!\rightarrow} \neg r)$	1, 3, 4, axiom 1

The proposition that if *p*, would *r* is stronger than the proposition that if *p*, might *r*. So this alternative derivation proceeds from stronger assumptions. But the derivation does not appeal to the duality of ‘might’ and ‘would’ counterfactuals.

To sum up: independently of various semantic assumptions, one can show that reverse Sobel sequence cases fit the conditions stated in (EI). ‘If *p* and *r*, would not-*q*’ raises a possibility to salience, and that possibility contradicts the proposition expressed by ‘if *p*, would *q*’. Given (EI), this entails that the speaker of the second sentence is epistemically irresponsible. My proposal is that the second sentence of a reverse Sobel sequence is infelicitous because it is an epistemically irresponsible thing to say.

So far I have taken possibilities to be propositions. Instead, you might say that a possibility is a world, and that a possibility is salient in the sense relevant to (EI) simply when it is contained in the context set of a conversation. (EI) would then entail that a speaker is epistemically irresponsible if she asserts a proposition that is false at some possibility in the context set of her conversation, if she cannot rule out that this possibility is actual. You might also say that possibilities are added to the context set as speakers accommodate presuppositions. For instance, you might say that in the zoo context described above, ‘cleverly disguised mules are not born with stripes’ presupposes that the caged animal in view might be a cleverly disguised mule. You might say that (3a) presupposes that it might be the case that if Sophie had gone to the parade, she might have been stuck behind a tall person. On this theory, a speaker should not utter (3b) or (5b) because she would thereby express a proposition false at live possibilities that are contained in the context set once she accommodates the presuppositions of (3a) or (5a).[16]

This presupposition-based theory shares a lot with the dynamic accounts discussed in §2. On all these theories, the first sentence of a reverse Sobel sequence affects the context by introducing a demand—roughly, a demand about some possibility—and this causes the second sentence to be infelicitous. But even this presupposition-based variation on my proposal differs from the dynamic accounts. There are technical differences: von Fintel and Gillies say that the trouble-making possibility is a world in which Sophie goes to the parade and gets stuck behind a tall person, whereas on the account just sketched, it is a world in which Sophie gets stuck behind a tall person *in some of the closest worlds in which she goes to the parade*.

Furthermore, I already mentioned a more significant difference: on all the dynamic accounts, the *truth* of the second sentence of a reverse Sobel sequence depends on what possibilities have been raised. The analogous claim about the zebra sequence would be that the truth of ‘that animal was born with stripes’ depends on whether someone has raised the possibility that the designated animal is a cleverly disguised mule. On my account of reverse Sobel sequences, what possibilities have been raised need not affect whether a Sobel sequence sentence is true, but only what a speaker must do to responsibly utter the sentence. This is a key difference between the dynamic accounts and the pragmatic theory developed in this section.

4 Arguments for my analysis

One reason to prefer my analysis to a dynamic semantics is that it is an independently motivated, more general theory. There must be some explanation for why the zebra sequence (5) is bad. Once we have developed (EI) to account for (5), we get an explanation for reverse Sobel sequences for free. Gillies and von Fintel, on the other hand, posit semantic rules specifically to account for the infelicity of sequences of counterfactuals. The rules are part of the lexicon. (EI) explains the same data by appealing to general, independently plausible facts about conversation and reasoning. So my analysis shares a general virtue of pragmatic theories: it explains more, using less.

Furthermore, unless Gillies and von Fintel augment their theories with a pragmatic account that is informed by my discussion in §3, my analysis more accurately predicts our judgments about a wide range of data. I said in §3 that certain sequences of counterfactuals are *generally* infelicitous, because generally the conditions in (EI) are met when they are uttered. But there are exceptions to these generalities. In exceptional cases, some conditions in (EI) will fail. It is straightforward for the pragmatic theory I have developed to yield correct predictions about these cases.

For example, my analysis naturally explains our intuitions about cases in which condition (iv) of (EI) fails. Remember that (3b) is generally infelicitous because speakers of (3b) are generally asserting propositions incompatible with salient possibilities that they cannot rule out. For instance, a speaker who utters (3b) after (3a) generally cannot rule out the possibility that Sophie might have been stuck behind a tall person if she'd gone to the parade. But these generalizations about speaker ignorance do not apply to every reverse Sobel sequence scenario. In some cases, a speaker who utters a reverse Sobel sequence may have some independent reason to utter the first sentence, a reason that would be in play even if she could rule out the trouble-making possibility made salient by that sentence. She may then utter the first sentence despite being able to rule out that possibility. In this kind of case, condition (iv) of (EI) may not hold for the second sentence in the sequence. So my analysis says that the second sentence of this sort of reverse Sobel sequence will not exhibit the sort of infelicity typical of reverse Sobel sequences.

And indeed, this is just what we find. 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:

(6a) If John had proposed to Mary and she had said yes, he would have been really happy.
(6b) But if John had proposed, he would have been really unhappy.

In this reverse Sobel sequence scenario, it is okay to utter (6b) after (6a). Here is why: I still have a reason to utter (6a), even if I can rule out the possibility that Mary might have married John if he had asked her. I may utter (6a) and (6b) precisely in order to get you to rule out that possibility, without breaking my promise to Mary. In this kind of case, condition (iv) does not hold for (6b), and so (EI) does not entail that my utterance of (6b) is irresponsible.

Just the same thing can happen when you ask me for two independent pieces of information. Suppose you want to know whether the act of proposing would have led to John being happy, and you also want to know whether Mary really would have been a good partner for John. So you ask me not only whether John would have been happy if he had proposed, but also whether he would have been happy if he had successfully proposed. In this scenario, even if I can rule out the possibility that Mary might have said yes if John had proposed, I still have an independent reason to utter (6a), namely to answer your second request for information. In this kind of scenario, condition (iv) does not hold for (6b), so uttering (6b) is not irresponsible. Hence my account correctly predicts that (6b) after (6a) will not exhibit the sort of infelicity exhibited by (3b) after (3a).

No part of the theory developed by Gillies and von Fintel distinguishes (6) from other reverse Sobel sequences. So they have trouble predicting our judgments about (6) unless they augment their theory in some way. (6a) expands the domain over which counterfactuals quantify, so that it includes some worlds in which John proposes to Mary and she says yes. Once this happens, no utterance of (6b) should be true. This simply contradicts our intuitions about (6b) as uttered in the context described above.

Condition (iv) of (EI) may also fail when it is a stable feature of the common ground that potentially trouble-making possibilities do not obtain. For instance, even in a philosophy classroom, we are used to ruling out the possibility that kangaroos might have had crutches if they had lacked tails. So even in a philosophy classroom, the following sequence (cf. LEWIS 1973) may be felicitous:

(7a) If kangaroos had lacked tails but had crutches, they would have had no trouble staying upright.
(7b) But if kangaroos had lacked tails, they would have toppled over.

Here again, a speaker who utters (7a) generally has some independent reason to utter this sentence, despite being able to rule out the possibility that kangaroos might have had crutches if they had lacked tails. In this kind of scenario, condition (iv) does not hold for (7b), and again, my account correctly predicts that the second sentence of a reverse Sobel sequence will not exhibit the sort of infelicity typical of reverse Sobel sequences.

So far I have discussed cases in which condition (iv) of (EI) fails. But my analysis also accounts for our intuitions about cases in which condition (iii) of (EI) fails. Consider the following reverse Sobel sequence, due to John Hawthorne:

(8a) If Sophie had gone to the parade and been shorter than she actually is, she would not have seen Pedro.
(8b) But if Sophie had gone to the parade, she would have seen Pedro.

It is easy and natural to raise the possibility that if Sophie had gone to the parade, she might have been stuck behind a tall person. It is less natural to raise the possibility that if she had gone to the parade, she might have been shorter than she actually is. Of course, it is *possible* to raise this possibility. If we have just been talking about whether parade vendors would profit from selling height-affecting drugs at large events, then it will be easier to raise the possibility that Sophie might have been shorter if she had gone to the parade. I take it that (8) is generally infelicitous in contexts like these. But in other contexts, we may willingly overlook worlds in which shorter counterparts of Sophie attend the parade. In these contexts, uttering (8a) does not suffice to raise the possibility that she might have been shorter if she had gone to the parade. In such contexts, my theory correctly predicts that (8) is felicitous.

Having seen two ways in which (EI) conditions can fail, we can now see the exact nature of the data to be explained. Our aim is not to explain why certain sequences of counterfactuals are infelicitous. Our aim is to explain why each sequence of counterfactuals is infelicitous as uttered in certain contexts. Even our original reverse Sobel sequence (3) can be felicitous. For instance, suppose you belong to a mafia organized to manipulate the exact movements of every tall person who attends a parade. If I ask you whether your mafia is conspiring to corner Sophie, you could still have a reason to tell me (3a), even if you can rule out the possibility that she might have been stuck behind a tall person if she had gone to the parade. My account correctly predicts that the typical infelicity of (3b) will not be present in these contexts.

Our judgments about reverse Sobel sequences are further complicated by the fact that speakers can signal whether (EI) conditions hold. For instance, simply in uttering the second sentence of a reverse Sobel sequence, a speaker may signal that she does not satisfy condition (iv) of (EI). Sending this signal is especially easy when her audience is not sure what she knows. Moreover, a speaker may strengthen this signal in a number of ways, e.g. by speaking assertively, adopting a condescending tone, or responsibly acknowledging that her assertion has contentious consequences. For example, the following sequence may end up sounding perfectly fine:

(9a) If Sophie had gone to the parade and been stuck behind a tall person, she would not have seen Pedro.
(9b) But hey, listen up—I am telling you: if she had gone, she would have seen him.

Speakers may also signal that they wish to ignore certain salient possibilities for purposes of conversation. Deliberately ignoring possibilities is sometimes signalled by a tone of impatience. It is also easier to deliberately ignore possibilities which are taken to be improbable:

(10a) If Sophie had gone to the parade and been stuck behind a tree, she would not have seen Pedro.
(10b) Oh, *come on*—if she'd gone, she would have seen Pedro.

Deliberately ignoring possibilities is a way of ruling them out of consideration. Of course, it may be in some sense irresponsible to deliberately ignore possibilities. (EI) concerns only one kind of epistemic irresponsibility: the kind that comes when a speaker neglects salient live possibilities. If you rule out salient possibilities by deliberately ignoring them, then as far as (EI) is concerned, you are not irresponsibly neglecting those possibilities. That is why (10b) is not generally infelicitous in the same way that (3b) is generally infelicitous.

Having considered several felicitous utterances of reverse Sobel sequences, I should clarify the nature of the foregoing arguments for my theory.[17] Strictly speaking, neither my own theory nor its dynamic rivals entail that any sequences of counterfactuals will be felicitous. Stating sufficient conditions for felicity is too much to ask of any analysis; semantic and pragmatic theories generally only ever state necessary conditions for felicity. This is for good reason: it would be nearly impossible for any semantic theory to state sufficient conditions for the felicity of utterances, since utterances may be infelicitous for a wide host of reasons, including violations of syntactic or phonological rules and innumerable subtle principles about the proper relation between an utterance and the prior discourse context. In light of these general limitations, some might cautiously prefer my theory only on the following grounds: my theory is silent with respect to the felicitous dialogues currently under consideration, while rival dynamic theories are still in need of some principled account of context shifting in order to avoid wrongly predicting that utterances will be semantically inconsistent and therefore infelicitous.

That being said, one may recognize a sense in which my theory outperforms its dynamic semantic rivals. Linguists often argue for a theory on the grounds that it predicts whether a sentence is felicitous or infelicitous, *once independent sources of infelicity are controlled for*. In other words, it is common practice to distinguish one source of infelicity from others; a theorist can then claim that a sufficient condition for infelicity characterizes a particular sort of linguistic badness, i.e. that the condition is a necessary and sufficient condition for a sentence to be infelicitous in a certain distinctive way. To quote just one example, BARKER 2000 argues that a proposed sufficient condition for infelicity explains why a description "will be felicitous... as long as other independent felicity conditions are satisfied (such as consistency with the common ground)" (23).[18] One methodology helpful for this purpose is the use of minimal pairs to isolate sources of infelicity.[19] The idea is simple: a linguist starts by comparing a number of infelicitous utterances with felicitous ones that resemble them in almost every respect. By controlling for possible sources of infelicity, she can develop a more precise theory of exactly what distinguishes the infelicitous utterances from the felicitous ones. For example, Chomsky's Binding Theory states only sufficient conditions for infelicity. But it is predictively powerful in virtue of mapping the exact contours of our judgments that sentences with certain arrangements of antecedents and pronouns are infelicitous while other very similar sentences are okay.

A single reverse Sobel sequence uttered in different contexts yields a useful minimal pair of utterances. By examining felicitous and infelicitous utterances of the very same sentence, we immediately rule out a number of potential sources of infelicity, including any syntactic features of the sentence and any dynamic semantic features that the sentence has independent of its context of utterance. It is a virtue of my theory that it fits the contours of our judgments about such minimal pairs. It is in this restricted sense that my theory accounts for our intuitions about felicitous and infelicitous reverse Sobel sequences. Strictly speaking, my claims are guarded: first, my theory does not need to resort to any context shifting in order to avoid making wrong predictions about the cases I have presented. Second, my theory characterizes the distinctive sort of linguistic badness often exhibited in reverse Sobel sequences, giving necessary and sufficient conditions for felicity when supplemented by the sort of *ceteris paribus* clause commonly understood to accompany such predictions.

Here is one more sort of argument in favor of my pragmatic account: my analysis not only accounts for our intuitions about felicitous reverse Sobel sequences; it also explains our intuitions about infelicitous counterfactuals in other linguistic contexts. Consider the following sequence:

(11a) Do you remember when Kate got stuck behind a tall person and missed seeing Pedro in her first baseball parade?
(11b) #But if Sophie had gone to the New York Mets parade, she would have seen Pedro.

(11a) is not a counterfactual. But it nevertheless raises the possibility that if Sophie had gone to the parade, she might have been stuck behind a tall person. My analysis predicts that (11b) is therefore infelicitous. Gillies and von Fintel do not predict this. Since (11a) is not a counterfactual, or even a modal sentence, it does not prompt any expansion of the domain over which counterfactuals quantify. So the dynamic semantic theory predicts that (11b) should sound as good as the second sentence of a Sobel sequence.

Unlike this argument concerning (11), most of the arguments given above are not sharply at odds with a strict conditional semantics. In order to clarify the limits of the present discussion, I would like to end with a proposal on behalf of the strict conditional theorist. One might claim that audiences adopt the following principle of charity: in cases where an audience believes a speaker is responsible, e.g. cases in which independent signals suggest that a sentence typically uttered by an irresponsible speaker is in fact uttered responsibly, one should attempt to interpret the second sentence of a reverse Sobel sequence as quantifying over a domain so that the counterfactual itself comes out true. The strict conditional theory could use such a principle to mimic the verdicts of the theory given in §3: whenever I claim that a reverse Sobel sequence is felicitous because the irresponsibility typical of speakers of such sequences is absent, the strict conditional theorist could claim that the signalled absence of such irresponsibility prompts a charitable contraction of the contextually supplied counterfactual domain. I leave it as an exercise for the reader to apply this sort of pragmatic theory to account for the felicity of (6)–(10) above.

In discussing the sort of sequences that one might investigate in order to give a theory of (6)–(10), VON FINTEL 2001 says "I do not have a theory of when and how the modal horizon can be expanded and contracted by expressions other than conditionals" (24), and GILLIES 2007 says "I have no good story to offer—indeed, no story at all—for how or why or in what cases such shrinkage is possible" (32). Gillies and von Fintel may see the present proposal as a friendly amendment: the foregoing reverse Sobel sequences help us understand the pragmatic theory that the strict conditional theorist should be looking for.[20] If they accept this olive branch, my main contention at present is simply that the investigation of pragmatic theories should lead us to recognize that the original reverse Sobel sequences that they discuss have no bearing on the question of whether we should abandon the standard semantics for counterfactuals. Doing justice to the full range of judgments about sequences of counterfactuals—whether in the context of a strict or variably strict conditional semantics—requires a pragmatic theory that is sensitive to myriad contextual factors not explored by von Fintel or Gillies. I have developed such a theory and demonstrated that it can indeed account for the original asymmetry between Sobel and reverse Sobel sequences.

5 'Might' counterfactuals

Our project is far from over. I will end with a few remarks about another potential application of the pragmatic approach: infelicitous sequences containing 'might' counterfactuals. Playing devil's advocate for the semantic approach, I will give one reason to think my §3 account does not fully explain why these sequences are infelicitous. But remaining optimistic about a pragmatic approach, I will state some desiderata for an analysis of 'might' sequences, and argue that the stated desiderata rule against some popular semantic accounts of their infelicity.

Consider the following sequence:

(12a) If Sophie had gone to the parade, she might have missed Pedro.
(12b) #But if Sophie had gone to the parade, she would have seen Pedro.

How should we explain the infelicity of (12b) after (12a)? Recall that Lewis accepts that 'might' and 'would' counterfactuals are duals. So Lewis could say that (12b) sounds bad because it is incompatible with (12a). However, we need not limit ourselves to a semantic explanation of the infelicity of

(12). Given the success of the pragmatic analysis so far, we might expect an alternative, pragmatic explanation of this infelicity.

However, extending the pragmatic approach to (12) is not straightforward. There is reason to think that (EI) does not fully explain why (12) is infelicitous. (12b) sounds bad when asserted after (12a). But in addition, (12) sounds bad in the context of a supposition:

(13) #Suppose that if Sophie had gone to the parade, she might have missed Pedro, but that if she had gone to the parade, she would have seen Pedro.

Something must explain why (13) is infelicitous. And (EI) alone cannot explain it. (EI) says that it is epistemically irresponsible to *express* a proposition incompatible with a salient live possibility. But that is not something that I do when I ask you to *suppose* that if Sophie had gone to the parade, she would have seen Pedro.[21]

To see this point another way, remember that the conditions of (EI) are generally met when a speaker utters (5b):

(5a) Cleverly disguised mules are not born with stripes.
(5b) #But that animal was born with stripes.

But it is perfectly okay to utter (5b) after (5a) in the context of a supposition:

(14) Suppose that cleverly disguised mules are not born with stripes, but that that animal was born with stripes.

The conditions of (EI) fail for (5), but it is okay to suppose (5). So it cannot be bad to suppose (12) only because the conditions of (EI) fail for (12). Something extra is wrong with (12), something that explains why it is infelicitous even in the context of a supposition.

In my view, (12) is not fundamentally different from a traditional reverse Sobel sequence: (12) is infelicitous for pragmatic reasons. Defending a pragmatic account would involve defending claims about the semantics of 'might' counterfactuals and about the behavior of various modals in the context of suppositions, and I will not undertake this project here. However, I will state two desiderata which make trouble for some semantic accounts of (12).

Desideratum one: an account of (12) should recognize similarities between (12) and (15):

(15a) Sophie might not see Pedro.
(15b) #But Sophie will see Pedro.

Note that like (12), (15) continues to be infelicitous in the context of a supposition:

(16) #Suppose that Sophie might not see Pedro, but that she will see Pedro.

Until faced with good reasons to give disparate explanations for these data involving 'would' counterfactuals and future contingents, it seems preferable to give a unified account of such similar phenomena. The same goes for conditionals embedding future contingents, such as:

(17a) If Sophie goes to the parade, she might not see Pedro.
(17b) #But if she goes to the parade, she will see Pedro.

Some theorists give a similar semantics for pairs of conditionals such as:

(3b) But if she had gone to the parade, she would have seen Pedro.

(17b) But if she goes to the parade, she will see Pedro.

If we accept a unified theory of "had-would" and "does-will" conditionals, there is even more pressure to find a unified explanation of the infelicity of (12) and (17). Ultimately, these sequences may not be infelicitous for exactly the same reason. But minimally, semanticists concerned with (12) should also be mindful of judgments about similar sequences containing future contingents.

Desideratum two: an account of (12) should explain the embedding behavior of (12b). For instance, (12b) is not felicitous after (12a), but neither is its negation:

(12a) If Sophie had gone to the parade, she might have missed Pedro.
(12b) #But if Sophie had gone to the parade, she would have seen Pedro.
(12c) Hey, look, you can't say that, because you don't know whether she would have seen him if she'd gone. / #Hey, look, you can't say that, because it's just false that she would have seen him if she'd gone.

It is not felicitous to deny (12b) when it is embedded in a question:

(18a) If Sophie had gone to the parade, she might have missed Pedro.
(18b) So would she have seen him if she'd gone?
(18c) I don't know for sure. / #No.

It is okay to assign (12b) a high subjective probability:

(19a) If Sophie had gone to the parade, she might have missed Pedro.
(19b) But I suspect that she would have seen him, if she'd gone.

In these sequences, there is a striking contrast between the embedding behavior of (12b) and sentences that are generally agreed to be false, such as (20):

(20) If Sophie were to go to the parade, she would definitely see Pedro.

Unlike (12b), the embedding behavior of (20) confirms that it is false:

(21a) If Sophie had gone to the parade, she might have missed Pedro.
(21b) #But if Sophie had gone to the parade, she would have definitely seen Pedro.
(21c) #Hey, look, you can't say that, because you don't know whether she would definitely have seen him if she'd gone. / Hey, look, you can't say that, because it's just false that she would definitely have seen him if she'd gone.

(22a) If Sophie had gone to the parade, she might have missed Pedro.
(22b) So would she definitely have seen him if she'd gone?
(22c) #I don't know for sure. / No.

(23a) If Sophie had gone to the parade, she might have missed Pedro.
(23b) #But I suspect she would definitely have seen him, if she'd gone.

The embedding behavior of (12b) suggests that (12b) is not straightforwardly false: it is more natural to negate and deny false utterances, and less natural to assign them high probability. Explanations of why (12b) is bad should at least accommodate these data, if not predict them.

These desiderata rule against an increasingly popular hypothesis about counterfactuals, recently defended in HÁJEK 2007. Hájek claims that 'might' counterfactuals like (12a) are almost always true:

(12a) If Sophie had gone to the parade, she might have missed Pedro.
(12b) #But if she had gone to the parade, she would have seen Pedro.

Hájek observes that (12) sounds contradictory. He concludes that (12a) and (12b) are contraries, and that (12b) is therefore false. One may repeat this argument for most counterfactuals. On these grounds, Hájek concludes that most 'would' counterfactuals are false.

In stating desiderata for a theory of (12), I have raised two worries for Hájek's argument. My first worry: one could use the same strategy to argue that most future contingents are false, on the grounds that sequences like (15) sound contradictory:

(15a) Sophie might not see Pedro.
(15b) #But Sophie will see Pedro.

But this would be an unwelcome conclusion. Some think that past utterances of future contingents had or have indeterminate truth values. But it is hard to accept that utterances of future contingents are automatically *false*. For instance, most theorists strongly resist saying my utterance yesterday of 'I will be alive tomorrow' is or was false, when I am in fact alive today.[22]

My second worry: consider again our judgments about (12c), (18), and (19):

(12c) Hey, look, you can't say [that if Sophie had gone to the parade, she would have seen Pedro], because you don't know whether she would have seen him if she'd gone. / #Hey, look, you can't say that, because it's just false that she would have seen him if she'd gone.

(18c) [Would Sophie have seen Pedro if she'd gone?] I don't know for sure. / #No.

(19) If Sophie had gone to the parade, she might have missed Pedro. But I suspect that she would have seen him, if she'd gone.

These judgments suggest that (12b) is not false. I think that Hájek might respond to this worry by saying that our judgments about (12c), (18), and (19) do not accurately signal whether (12b) is false. Hájek says that our practice of uttering counterfactuals such as (12b) is "legitimated" by the existence of "nearby" true counterfactuals such as (24):

(12b) #If Sophie had gone to the parade, she would have seen Pedro.

(24) If Sophie had gone to the parade, she would very probably have seen Pedro.

Hájek says that since (24) is true, and closely related to (12b), we may legitimately assert the latter:

> There are true counterfactuals closely related to the ones we assert that support our practice, at least when the prevailing standards for asserting counterfactuals are somewhat forgiving, as they typically are on the street. So we can legitimately assert various counterfactuals. Still, most of them remain false. (52)

Hájek could go on to say that since (24) is closely related to (12b), we may legitimately judge embedded occurrences of (12b) as if they were occurrences of (24). Moreover, we do in fact judge embedded occurrences of (12b) in this way. So our judgments about (12c), (18), and (19) reflect whether (24) is false, not whether (12b) is false.

To respond: we do not in fact judge embedded occurrences of (12b) as if they were occurrences of (24). For instance, compare the following sequences, as uttered by speakers who know that Sophie is an extremely tall and aggressive Pedro fan:

(19a) If Sophie had gone to the parade, she might have missed Pedro.
(19b) So would she have seen him if she'd gone?
(19c) I don't know for sure.

(25a) If Sophie had gone to the parade, she might have missed Pedro.
(25b) So would she have very probably seen him if she'd gone?
(25c) #I don't know for sure.

It is more natural to attribute knowledge of propositions expressed by counterfactuals that embed a 'would probably' operator as (24) does, and it is harder to attribute knowledge of propositions expressed by ordinary 'would' counterfactuals such as (12b). But our judgments about the 'would' counterfactual sequence (19) and the 'would probably' sequence (25) distinguish (12b) from (24). So it is reasonable to assume that our judgments about embedded occurrences of 'would' counterfactuals in (12c), (18), and (19) suggest that it is the 'would' counterfactual (12b) itself, not just its 'would probably' counterpart (24), that is not false. This leaves open several explanations of why the ordinary counterfactual (12b) is infelicitous. But it does tell against semantic explanations of the sort that Hájek gives.[23]

Notes

[1] See Gillies 2007 for a recent discussion of the relevant literature.

[2] When I am sure it will not cause any confusion, I will cut corners to make claims more readable, e.g. where '*p*' is a schematic letter to be replaced by a sentence, I use '*p* worlds' to refer to worlds where the semantic value of that sentence as uttered in the understood context is true.

[3] Lewis 1973 thanks J. Howard Sobel for bringing sequences like (2) to his attention.

[4] More precisely, Lewis 1973 says that 'if *p*, would *q*' expresses a proposition that is non-vacuously true at a world just in case some *p*-and-*q* world is closer to that world than any *p*-and-not-*q* world. Stalnaker 1968 says that 'if *p*, would *q*' expresses a proposition that is true at a world just in case the closest *p* world is a *q* world. I will not focus on the details of these rival versions of the standard semantics, but I will flag differences between the accounts where they are relevant to my arguments.

[5] In (2001), von Fintel credits Irene Heim with the origination of reverse Sobel sequences.

[6] Other proponents of a return to the strict conditional analysis include Warmbrod 1981 and Lowe 1995.

[7] Strictly speaking, the accessibility function maps each world to its own domain of accessible worlds. The truth of a counterfactual at a world depends on properties of the worlds accessible from that world. So in a sense, a counterfactual quantifies over many domains: one for each world. I trust the reader to read my claims accordingly.

[8] What makes a semantic theory dynamic is controversial. Some may prefer to reserve the term 'dynamic' for the first version of von Fintel's proposal. I follow Gillies 2007 in applying 'dynamic' to theories resembling either version.

[9] I discuss a sequence about zebras because zebra examples are familiar, and so using a zebra example is a quick and reliable way to situate my theory among familiar debates. However, our familiarity with zebra examples can create unwanted noise in our judgments about them. Some informants judge there to be a more marked difference between the following conversations, as uttered in New York City:

(4a′) My car is around the corner.
(4b′) But cars get stolen in New York City all the time.

(5a′) Cars get stolen in New York City all the time.
(5b′) #But my car is around the corner.

[10] One might aim to derive this principle from others, e.g. from the knowledge norm of assertion and the principle that a speaker knows a proposition only if she can rule out salient possibilities incompatible with that proposition. But (EI) must eventually follow from norms more general than those governing assertion, since some analog of (EI) must explain why it can be irresponsible even just to *think* a reverse Sobel sequence.

[11] Here I am taking possibilities to be propositions. Strictly speaking, one could endorse a family of precisifications of (EI) corresponding to various necessary and sufficient conditions for "ruling out" a proposition. For sake of simplicity, I will take it that the reader has intuitions informed by his or her understanding of this term of ordinary language; my argument references these intuitions.

[12] For simplicity, I talk as if infelicity is a property of utterances. Strictly speaking, infelicity is audience-relative: an utterance sounds infelicitous to an agent insofar as she takes the resulting assertion to be epistemically irresponsible.

[13] See Lewis 1973, p. 132 for a complete axiomatization for **VC**.

[14] This means that the felicity conditions outlined by (EI) are sensitive to what is common ground when an assertion is made. (EI) is comparable with other pragmatic norms in this respect. For example, the *informativeness* of an utterance—in the sense of the quantity maxims in Grice 1987—depends on what is common ground when the utterance is made.

[15] Lewis 1973 demonstrates that the duality thesis and the law of conditional excluded middle together entail the equivalence of 'might' and 'would' counterfactuals. Many have responded to this argument by rejecting the duality thesis; see Williams 2009, DeRose 1997, 1994, and Heller 1995 for some examples. See Stalnaker 1981 for arguments in favor of the law of conditional excluded middle.

[16] I do not endorse this theory. Presuppositions are essentially marked by the way they project through some environments and not others, yet (3a) and (5a) make the relevant possibilities salient regardless of what linguistic environments they are in. But this presupposition-based theory has a lot in common with my proposal, so it is instructive to contrast this theory with the dynamic accounts, to highlight differences between the dynamic accounts on the one hand, and the presupposition-based theory and my proposal on the other.

[17] Thanks to Alan Hájek for prompting the present discussion of linguistic methodology.

[18] For other examples of similar claims, see the account of felicitous indefinite determiners in Alonso-Ovalle et al. 2009 and the "evidence-based account of the felicity conditions" of disjunctive sentences in Simons 1999.

[19] See Fitzgerald 2009 for reflective remarks about using the "'minimal pair' experimental setting" to control for causes of infelicity not of experimental interest (20).

[20] See Moss 2010 for a more detailed discussion of concerns that distinguish this strict conditional account from my pragmatic theory, and for arguments that such concerns ultimately favor my analysis.

[21] (EI) may have more general analogs that do not merely govern assertion. But since many audiences suggest that it is significantly more felicitous to suppose than to assert reverse Sobel sequences, I do not endorse an analog of (EI) for supposition.

[22] This problem is not devastating. Hájek simply owes us some explanation for why his argumentative strategy does not overgenerate. I expect that his explanation will appeal to apparent disanalogies in our judgments about counterfactuals and future contingents; for instance, truth values of future contingents may supervene on facts about the actual world while truth values of counterfactuals arguably cannot. I will not address various potential arguments here, except to say that the abovementioned judgments about embedded counterfactuals suggest that ordinary language judgments about 'would' counterfactuals bear more similarities to judgments about future contingents than is commonly recognized.

[23] Thanks especially to Alan Hájek, Kai von Fintel, Thony Gillies, Ofra Magidor, Bob Stalnaker, Eric Swanson, and audiences at NYU, Berkeley, Michigan, Rutgers, and BSPC 2008 for helpful comments.

References

Alonso-Ovalle, Luis, Paula Menéndez-Benito & Florian Schwarz. 2009. "Maximize Presupposition and Two Types of Definite Competitors." *North East Linguistic Society (NELS)*, vol. 39.

Barker, Chris. 2000. "Definite Possessives and Discourse Novelty." *Theoretical Linguistics*, vol. 26 (3): 211–228.

DeRose, Keith. 1994. "Lewis on 'Might' and 'Would' Counterfactual Conditionals." *Canadian Journal of Philosophy*, vol. 24 (3): 413–418.

———. 1997. "Can It Be That It Would Have Been Even Though It Might Not Have Been?" In *Philosophical Perspectives 11: Mind, Causation, and World*, James Tomberlin, editor, 385–413. Blackwell Publishers, Ltd., Oxford.

von Fintel, Kai. 2001. "Counterfactuals in a Dynamic Context." In *Ken Hale: A Life in Language*, Michael Kenstowicz, editor. MIT Press, Cambridge.

Fitzgerald, Gareth. 2009. "Linguistic Intuitions." The *British Journal for the Philosophy of Science*, vol. 61 (1): 123–160.

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

Grice, Paul. 1987. "Logic and Conversation." In *Studies in the Way of Words*. Harvard University Press, Cambridge.

Hájek, Alan. 2007. "Most Counterfactuals are False." Ms., Australian National University: http://philrsss.anu.edu.au/people-defaults/alanh/paper s/MCF.pdf.

Harper, William L., Robert Stalnaker & Glenn Pearce, editors. 1981. *Ifs: Conditionals, Belief, Decision, Chance, and Time*. D. Reidel Publishing Company, Dordrecht.

Heller, Mark. 1995. "Might-Counterfactuals and Gratuitous Differences." *Australasian Journal of Philosophy*, vol. 73: 91–101.

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

Lowe, E. J. 1995. "The Truth about Counterfactuals." *Philosophical Quarterly*, vol. 45: 41–59.

Moss, Sarah. 2010. "Constraining Credences in Counterfactuals." Ms., Department of Philosophy, University of Michigan.

Simons, Mandy. 1999. "On the Felicity Conditions of Disjunctive Sentences." *Proceedings of the Western Conference on Linguistics*, vol. 11.

Stalnaker, Robert C. 1968. "A Theory of Conditionals." In Harper et al. (1981), 41–55.

———. 1978. "A Defense of Conditional Excluded Middle." In Harper et al. (1981), 87–104.

———. 2002. "Common Ground." *Linguistics and Philosophy*, vol. 25: 701–721.

Warmbrod, Ken. 1981. "Counterfactuals and Substitution of Equivalent Antecedents." *Journal of Philosophical Logic*, vol. 10: 267–289.

Williams, Robbie. 2008. "Conversation and Conditionals." *Philosophical Studies*, vol. 138 (2): 211–223.

———. 2009. "Defending Conditional Excluded Middle." Forthcoming in *Noûs*.