"Now that you mention it, I wonder..." : Awareness, attention, assumption

de Jager, S.T.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)

Download date: 11 Dec 2018
Chapter 4

Case study: Sobel sequences

“Apparently there is no limit,” Joe remarked.
“Anything can be said in this place and it will be true and will have to be believed.”

Flann O’Brien, The Third Policeman

Since [Sta68] and [Lew73], a similarity-based account of counterfactual conditionals has been pretty much standard in philosophical logic. According to such a theory (in very broad brush-strokes), a counterfactual conditional as in (1) is true in a world $w$ if in all those worlds where $p$ holds which are most similar to $w$, $q$ holds also.

\begin{equation}
\text{(1) If } p \text{ were the case, } q \text{ would have been. \textbf{[Notation: } p \sqsupset q]} \\
\begin{align*}
a. & \text{ If Sophie had gone to the New York Mets parade, she would have seen Pedro Martínez.} \\
b. & \text{ If the US threw all its nuclear weapons into the sea tomorrow, there would be war.}
\end{align*}
\end{equation}

Two significant problems exist with this account, which this chapter will address. The first concerns the notion of ‘similarity’: if this cannot be given a systematic foundation, then a semantics based on similarity cannot be truly explanatory. The second problem concerns the data: some combinations of counterfactual sentences (‘Sobel sequences’) show context-dependent dynamic behaviour which cannot be explained on the static account, and the most influential solutions for this problem have advocated giving up the similarity-based account almost entirely.

I shall take these problems in reverse order. My first aim is to explain the dynamic behaviour of Sobel sequences as a minimal addition to a similarity-based account, adding nothing but awareness dynamics to the picture. The explanation will provide quite strong constraints on the content of ‘similarity’, which fit in neatly with the most recent attempts to explicate ‘similarity’ in causal terms.
Chapter 4 · Case study: Sobel sequences

**Definition 4.1:** Object language with counterfactuals $\mathcal{L}^{\Diamond\to}$. The object language with counterfactuals adds a binary counterfactual conditional operator $\Diamond\to$ to the syntax of $\mathcal{L}^\Diamond$.

Recall that $\mathcal{L}^\Diamond$ is the propositional language with might. As I said when introducing the first object languages, I prefer to leave the syntax relatively loosely constrained: $(\text{might } p) \Diamond\to (q \Diamond\to (\text{might } r))$ is strictly speaking a well-formed formula. Most nestings will not make sense with the semantics I give, but might in the consequent of a counterfactual is deliberately included.

1. **Counterfactual dynamics: the problem**

Since I want to defer the exploration of just what ‘similarity’ should mean to Section 3, let me follow [Lew81] and refer to **ordering semantics**; indeed, I will largely follow his presentation of ordering semantics in that paper.

**Definition 4.2:** Ordering semantics (after [Lew81]). Let $W$ be a finite set of worlds. An **ordering frame** for $W$ associates with each world $w$ a preorder $\leq_w$ on $W$ (that is, a reflexive and transitive binary relation $\leq_w \subseteq W \times W$), satisfying the following condition:

**Centering:** For all $w, v \in W$: if $v \neq w$ then $w <_w v$.\(^1\)

We write $S^W$ for such an ordering frame ($S$ for “similarity”); it is a function from elements of $W$ to orderings on $W$, but we write $\leq_w$ for the ordering associated with $w \in W$, instead of the more cumbersome $S^W(w)$.

Given any preorder $\leq$ on $W$, the set of $\leq$-**minimal** $\varphi$-worlds is written

$$\text{min}_{\leq}(W \upharpoonright \varphi)$$

and defined

$$\{w \in W ; w \models \varphi \land \neg \exists w' \in W : w' \models \varphi \land w' < w\}.$$  

(The finite setting guarantees the well-behaved existence of such sets.) The set of closest $\varphi$-worlds to $w$ is given by

$$\text{min}_{\leq_w}(W \upharpoonright \varphi).$$

The counterfactual $\varphi \Diamond\to \psi$ is true at world $w$ if all closest $\varphi$-worlds to $w$ are $\psi$-worlds:

$$w \models \varphi \Diamond\to \psi \iff \forall w' \in \text{min}_{\leq_w}(W \upharpoonright \varphi) : w' \models \psi.$$  

The ordering semantics stands in opposition particularly to a **strict** analysis (more standardly assumed for indicative conditionals), under which $p \Diamond\to q$ would be true if $q$ holds in all the worlds where $p$ does. (The set of possible worlds must be contextually restricted in some fashion, but according to the

\(^1\)I assume Lewis’s constraint “Universality”: that all worlds are included in the field of $\leq_w$.}
strict analysis this restriction is not directly related to the semantics of the counterfactual.) The strict analysis validates the inference pattern ‘strengthening the antecedent’: from $p \implies q$ it follows that $p \land r \implies q$. That counterfactual conditionals do not allow strengthening the antecedent was a strong motivation for the ordering semantics; here is an example, which we will see reappear in many variations throughout the rest of the chapter.\(^2\)

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

An ordering semantics allows these two counterfactuals to be simultaneously true at a world $w$, if the ordering relation is suitably chosen. However something more seems to be going on.\(^{[Fin01]}\) credits Irene Heim with the observation that in reverse order this sequence no longer sounds felicitous:

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

Notation 4.3: Some informal terminology. A Sobel sequence is a pair of counterfactuals with the forms shown in (2) and (3). The forward sequence is as in (2), the reverse sequence as in (3). Based on the number of distinct possibilities alluded to, I call (2-a) and (3-b) simple, and (2-b) and (3-a) complex.

The Sophie Sobel sequence is by no means an isolated example — perhaps even the majority of ‘natural’ examples of the failure of strengthening the antecedent behave in this way. (It turns out, however, that the pattern is not universal; I will give examples below that share the same logical structure but for which both orderings are felicitous.) The challenge to the ordering semantics is clear: as a static theory it cannot represent the influence of ordering in discourse, so there is no chance for it to explain these data.

I will argue for a modular approach: the core of the ordering semantics should be left exactly as it is, but combined with a separate component dealing with dynamic effects; I want to argue that the context-change in these examples is properly modelled as an awareness update (and thus has nothing inherently to do with counterfactuals at all).

This account has three benefits: it lets us keep the familiar ordering semantics, with all its acknowledged benefits; since the mechanism of awareness

---

\(^2\)I have the example from \([Mos07]\); \([Gil07]\) has “If Sophie had gone to the parade, she would have seen Pedro dance.” According to Wikipedia, Pedro Martínez was a pitcher for the New York Mets from 2005 until 2008 and is now a free agent; I don’t know whether he typically danced at parades.
update is not associated with counterfactuals, we can reuse the same explanation for a large number of other constructions that pattern like Sobel sequences (Section 4); and the specific requirements of this account for the ordering semantics will throw some light on what ‘similarity’ should mean (Section 3).

First, though, I have to discuss an alternative explanation for the contrast between (2) and (3), which has recently grown in popularity. The reader more interested in my positive contribution is invited to skip ahead to Section 2.

1.1 · Strict semantics and shifting context

At first sight the alternative is extremely seductive, especially in its more elaborate formulations. [War81] is the earliest version I have seen, but [Low90; Fin01; Gil07; Wil08] (among others) all give variants of this proposal, and I will cheerfully mix their terminology to suit myself. The version I give is very simple, and I must ask the reader to accept (or to check for themselves) that the bells and whistles added by the various accounts that I conflate will not affect my critique. Here, then, is the proposal in a nutshell; I call it the shifting strict analysis.

We are to evaluate a counterfactual according to the strict semantics, against a contextually given set of worlds, the modal horizon of [Fin01]. A counterfactual \( \varphi \square \rightarrow \psi \) comes with what [Gil07] calls an entertainability presupposition, that \( \varphi \) be satisfiable within the modal horizon. If the presupposition is not met, accommodation adds some worlds to the modal horizon; the worlds added are the closest \( \varphi \)-worlds to the evaluation world, according to the same kind of similarity ordering as is needed for ordering semantics. After accommodation, the strict truth conditions are simply: \( \varphi \square \rightarrow \psi \) is true (at \( w \)) if every world in the modal horizon (of \( w \)) which supports \( \varphi \) also supports \( \psi \).

It is easy to see that such an account explains the Sobel data. Entertainability presuppositions only add worlds to the modal horizon, so counterfactuals later in the discourse ‘inherit’ the possibilities introduced by earlier utterances (such as Sophie being stuck behind someone tall). More subtly, it can provide exactly the same predictions about single counterfactuals (uttered in the ‘null context’) as does the ordering semantics: if the modal horizon in the null context is

\[3\] Warmbröd was concerned with the inference pattern of substitution of equivalent antecedents rather than strengthening the antecedent; his proposal suffers in readability, through no fault of its own, by predating modern notions of dynamic semantics. Williams is concerned with indicative, rather than counterfactual, conditionals; we will consider some of his data below, in extending the account beyond counterfactuals. The account I give is based most directly on the von Fintel analysis. My impression is this has been relatively influential, which I find somewhat surprising in light of the simple counterexamples that I will introduce below. Sarah Moss is a welcome voice of scepticism, and I will draw heavily on her —as yet unpublished— account [Mos07] in what follows.

\[4\] I don’t think it is essential to these accounts that entertainability ‘presuppositions’ be presuppositions as usually conceived (in fact there are both systematic similarities and systematic differences). Using the term will help the clarity of my account, if I may do so without taking it too seriously.
suitably trivial (empty or containing only the evaluation world), the first update will add to it precisely those worlds from the ordering that would influence the truth conditions according to the ordering semantics.

1.2 · Problems with the shifting strict analysis

The way the shifting strict analysis incorporates single counterfactuals marks a profound difference in methodology from my own approach. Both analyses must acknowledge the fact that the ordering semantics provides extremely intuitive predictions for single counterfactuals (in the ‘null context’). My approach is to augment the ordering semantics with a component dealing with context change; the predictions in the null context still come from the same core mechanism in the ordering semantics. The shifting strict approach, on the other hand, supplants the ordering semantics; the predictions in the null context come from a different mechanism, which is carefully adjusted to produce coinciding predictions. This produces an odd redundancy in the system: the strict semantics is for counterfactuals, but the shiftiness is for counterfactuals too; the theory in some sense contains intertwined both a static and a dynamic counterfactual semantics.

I think this methodological distinction already provides reason to prefer my analysis, if the two can be shown to perform equally well on the data: I prefer a ‘modular’ approach in which the dynamic behaviour of counterfactuals emerges from interaction between a counterfactual semantics and a dynamic theory of discourse, where the two are much more distinct and independent than the shifting strict analysis will allow. I will go further, though, and suggest that there would be good reason to reject the shifting strict analysis even if we had no better account of the dynamics.

I want to raise three problems, respectively observational, conceptual, and methodological: the shifting strict analysis mispredicts on some simple examples, close to the core of what it is designed to explain; it rejects without good reason the possibility of simultaneously considering different ‘levels of counterfactuality’; and it fails to capture the generality of the Sobel forward-and-reverse pattern, by limiting its account to (at best) conditional sentences.

1.2.1 · Prediction failures

Each of the following sequences is predicted by the shifting strict analysis to be infelicitous, in the same way that the reverse Sobel sequence is. (I have (4-c) from [Mos07], where it is credited to John Hawthorne.) To my ear, at least, none of them bear out that prediction.

(4) a. (i) If Sophie had gone to the parade and not seen Pedro, she wouldn’t have seen Pedro. 
\[ p \land \neg q \rightarrow \neg q \]

(ii) But if Sophie had gone to the parade, she would have seen Pedro. 
\[ p \rightarrow q \]
Chapter 4 · Case study: Sobel sequences

b. (i) If Sophie had gone to the parade and not seen Pedro, she would have been upset. \[ p \land \neg q \rightarrow s \]
(ii) But if Sophie had gone to the parade, she would have seen Pedro. \[ p \rightarrow q \]
c. (i) If Sophie had gone to the parade and been shorter than she actually is, she would not have seen Pedro.
(ii) But if Sophie had gone to the parade, she would have seen Pedro.
d. (i) If Sophie had gone to the parade and been eaten by a dinosaur, she wouldn’t have seen Pedro. \[ p \land d \rightarrow \neg q \]
(ii) But if Sophie had gone to the parade, she would have seen Pedro. \[ p \rightarrow q \]

Of course any theory can deal with a limited number of counterexamples: there is always the possibility that further wheels upon wheels may be added that will save the phenomena. How attractive such a strategy appears should depend on how central the counterexamples are for the class of phenomena to be explained; my impression of these examples is that they are paradigm cases of counterfactual use, and that a semantics for counterfactuals should explain such data in its core rather than at the level of a corrective epicycle.

The four examples divide naturally into two classes. The first two involve entertaining the possibility that Sophie does not see Pedro, but (unlike the classic Sobel sequence) provide no reason to take this possibility seriously. The last two examples have a different flavour: they provide reasons why Sophie might not see Pedro, but the reasons themselves are not to be taken seriously.

All these examples can be modelled using orderings, but I want to make a stronger claim. In cases like the first two, where we have no reason to expect Sophie to not see Pedro, I only want to allow orderings in which she at least might see him (if she goes to the parade, naturally). Section 3 gives an account of where the orderings of ordering semantics come from, which fulfils this desideratum.

For the second two cases, it is enough for me that ordering semantics makes it possible that they be acceptable, since the shifting strict account does not. This leads us to the conceptual critique of the shifting strict analysis.

1.2.2 · Simultaneity ruled out

Under the ordering semantics the counterfactuals making up a Sobel pair can be simultaneously true. In the cases Heim noticed this doesn’t seem right; the impression we have on reading the forward sequence is that in light of the complex utterance, the simple is no longer true. However examples such as (4-d) above should make us question whether this pattern is (as the shifting

\[ ^5 \text{That the acceptability of (4-a) does not depend on its tautological status is shown by (4-b).} \]
strict analysis would have it) truly universal. It seems perfectly reasonable to believe simultaneously that Sophie would not have been eaten by a dinosaur at the parade, and nonetheless that from inside a dinosaur she would not have been able to see Pedro.

We might think of this example in terms of embeddings of counterfactual contexts. At the factual level, Sophie doesn’t go to the parade. In the first counterfactual context, she does go and sees Pedro — and is of course not interfered with by dinosaurs, which live only in a second counterfactual context, embedded in the first.

A counterfactual is most natural in a context where its antecedent is known to be false; we can agree at least on this, without taking a stand on precisely where this naturalness comes from (presupposition, Gricean inference, or whatever your favourite explanation may be). We can see this as a kind of concession: “I admit that in our current context \( \phi \) is the case, but if we go to an embedded context where it isn’t then. . . .” And this works equally well for higher levels: “I admit that in the embedded context where Sophie goes to the parade she is not eaten by a dinosaur, but if we go to yet further embedded context where she is then. . . .”

The odd thing about the shifting strict analysis is that it treats these two cases as fundamentally different. In the first case, what is ‘conceded’ is sacrosanct and cannot be affected by the counterfactual utterance; but in the second case, the concession is undermined by the effect of accommodating the entertainability presupposition. After “I agree that she didn’t, but suppose that she had”, “she didn’t” is still supported; after “I agree that she wouldn’t have, but suppose that she did”, “she wouldn’t have” is not.

Here is a similar, more extended example.

(5)  
\begin{enumerate}
  \item a. A: Suppose Sophie had gone to the parade yesterday.
  \item b. B: She would have seen Pedro.
  \item c. A: But suppose she was eaten by a dinosaur.
  \item d. B: She wouldn’t have been!
  \item e. A: Sure. But suppose she was. Then she wouldn’t have seen Pedro, right?
  \item f. B: Alright.
  \item g. A: So if she had gone to the parade, would she have seen Pedro?
  \item h. B: Of course she would have.
\end{enumerate}

I don’t want to claim that “Suppose that \( \phi \) were true; then \( \psi \) would be” always has a meaning identical to “If \( \phi \) were true \( \psi \) would be”, but the intuitive similarity should not be dismissed either. Intuitively we distinguish different nested counterfactual contexts in these examples, and so those distinctions should be available for our counterfactual semantics to work with.
1.2.3 · Missing generalisation

The final significant deficiency of the shifting strict analysis is that it has too narrow a focus and too specific a mechanism. The analysis posits particular features of the semantics of counterfactuals underpinning the Sobel pattern; if the same pattern occurs without involving counterfactuals at all, the shifting strict account has nothing to say about it. Sarah Moss has collected a wide range of non-conditional examples showing the Sobel pattern (as in (6)), and gives an intuitive explanation of their common structure in an unpublished manuscript [Mos07]; her explanation requires no revision of standard counterfactual semantics.

(6) a. (i) A: My car is around the corner.
   (ii) B: Cars get stolen all the time here in New York City.

b. (i) B: Cars get stolen all the time here in New York City.
   (ii) A: ? My car is around the corner.

Her observation is that attention to possibilities affects the assertability of perfectly ordinary statements as well as counterfactuals. I discuss some more of her data in Section 4, which extends my own theory beyond counterfactuals, since I disagree slightly on its interpretation, but I am far more sympathetic to her mode of explanation than to that of the shifting strict analysis.

1.3 · Desiderata for a replacement theory

I have dwelt on the shifting strict analysis at some length, because its shortcomings outline a number of desiderata that must be achieved if any theory is to qualify as a potential replacement.

Most obviously, we need a dynamic account. The dynamics cannot be driven by conditional form (if we are to cover Moss’s data); our theory should not predict that every sequence with the Sobel form has the Sobel pattern of acceptability (to cover the core counterexamples of (5)). We must incorporate enough of the ordering semantics to be able to represent nested levels of counterfactual supposition.

However, Heim’s observation and the widespread acceptance of (something like) the shifting strict analysis points at another, rather more subtle, desideratum. Sobel-pattern pairs are easy to think up, but pairs that are felicitous when reversed are quite a bit thinner on the ground. The shifting strict analysis rules such pairs out entirely, and falsely; but our theory should, ideally, have

---

6In fact not quite every account that I gather under the heading “the shifting strict analysis” is concerned solely with counterfactuals. The account of [Wil08] applies to indicative conditionals rather than counterfactuals, and is in some respects closer to my own views. It still ties the mechanism to the specific semantics of conditionals, though (entertainability presuppositions are triggered only by conditional antecedents); I am arguing for a general framework of awareness-sensitive semantics which applies equally to conditionals and to any other construction.
something to say about the relative frequency (and naturalness) of the two kinds of examples.

I will now give a theory that fulfils these desiderata.

2 · Orderings with awareness

The idea is simple in the extreme. We will simply add an ordering frame to the dynamic awareness models of the previous chapter, as a component of the model of reality; a counterfactual conditional will be evaluated according to the ordering semantics but on the awareness state of the agent, which may exclude a number of possibilities due to the agent’s assumptions. Changes in awareness will update this state as I described in the previous chapter; it is these dynamics, rather than anything particular to counterfactuals, which will give rise to the pattern noted by Heim.

The ordering semantics I gave in Definition 4.2 was for a single world, so our first task is to lift it to an information state. It might seem that a state should support the counterfactual if all worlds in the state do so, but the possibility of might occurring in the consequent requires a more careful approach. If I don’t know whether it is raining in Whitechapel or not, I can perfectly acceptably tell you “If you were in Whitechapel you might have been rained on”; presumably this is true even though one of my epistemic alternatives (where it is not raining in Whitechapel) does not itself support the counterfactual. [Vel05] has a similar starting point (although the semantics end up quite different). Veltman defines the result of updating an information state with $ϕ \rightarrow$ (something like “Suppose that $ϕ$”); the new state collects the nearest $ϕ$-neighbours of all the worlds in the state (call that set $C$, for “counterfactual possibilities”), and the consequent of the conditional is tested against this set as a whole.

This means that a counterfactual cannot carry information: like any test, it either leaves the information state unchanged or takes it to the absurd state. In Section 4.2 I give an alternative definition which allows counterfactuals to be informative ([Vel05] also gives both a test and an update version of his semantics), and some reasons to suppose that we need both definitions.

**Definition 4.4:** Ordering semantics for attention models (counterfactuals as tests). Let $M = \langle W, Ω, V, S^W \rangle$ be a model of reality augmented with an ordering frame, and $σ = \langle A, B, Ξ \rangle$ a state. We treat a counterfactual as a test:7

$$σ[ϕ \rightarrow ψ]_b^M = \begin{cases} σ & \text{if } \langle A, C, Ξ \rangle \models^M ψ, \\ \langle A, \emptyset, Ξ \rangle & \text{otherwise,} \end{cases}$$

where $C = \bigcup_{w \in B} \min A \upharpoonright ϕ$

---

7This definition formally allows for nesting counterfactuals; I make no claim for its appropriateness in such cases, however.
Note that this definition gives only the pure belief update \([\cdot]_b\). The key to the proposal is that the update \([\varphi \rightarrow \psi]_b\) is, as in the previous chapter, analysed in two steps: \([\varphi \rightarrow \psi]_n[\varphi \rightarrow \psi]_b\). The first step works just as any other attention update: it draws attention to atomic formulae mentioned in the counterfactual. It is only the second step that needs an explicit definition for the counterfactual connective.

For purely propositional \(\varphi\) and \(\psi\) the definition I give comes to the same as checking whether each world in \(B\) individually satisfies \(\varphi \rightarrow \psi\) with the standard world-based ordering semantics for the counterfactual, however when \textit{might} appears in the consequent the results can differ.

(7) The barroom tough Big Joe has invited you outside for fisticuffs, an invitation you have politely declined. You are not sure if Big Joe is drunk or not; sober he is unstoppable, but drunk he is as likely to knock himself out as to down his opponent. Describing the events the following day, you say:

a. “If I had taken the fight, I might have won.”

The semantics given above correctly predicts (7-a) to be true in this context; the world-based formulation with universal quantification, on the other hand, would make it false. (There is an epistemic alternative, namely the one in which Joe is sober, for which all closest fight-alternatives have you losing.)

Using such a semantics any Sobel-pattern pair of counterfactuals can be made simultaneously true, or incompatible, as required (the ordering semantics is all that is needed). However in order for the forward and reverse judgements to come out right under the awareness update, we need a particular structure on the ordering relation. Let me give the representation of Sophie’s visit to the parade first as a ‘Just So story’, and show how it matches the data. After that I can say what the account requires in general, and how orderings satisfying those constraints might be generated. Figure 4.1 on the facing page gives the worlds and similarity relation we will need.

We assume for the moment that the agent holds only \(w_0\) possible (\(B = \{w_0\}\)), so we are only concerned with the ordering for that world. Most of the details of the ordering should be uncontroversial; the only element that is both necessary for my account and potentially unexpected is the equisimilarity of \(w_5\) and \(w_6\). In fact it is only required for my account that \(w_5\) be not more similar to \(w_0\) than

\^{8}Note that this discussion is independent of the well-known debate on whether the law of conditional excluded middle holds at a particular world. If it does, then there are three relevant epistemic alternatives: one in which Joe is sober, and two in which he is drunk (differing only in who wins in the closest alternative in which you fight). I prefer a formulation in which the uncertainty about the outcome of the fight is metaphysical rather than epistemic, which I will argue for in Section 3. In either case, however, the lifting to information states provides counterexamples to conditional excluded middle — just as an information state may record neither certainty that \(p\) nor that \(\neg p\).
Orderings with awareness

<table>
<thead>
<tr>
<th></th>
<th>$w_0$</th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>$w_3$</th>
<th>$w_4$</th>
<th>$w_5$</th>
<th>$w_6$</th>
<th>$w_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$q$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$r$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$w_0 < w_0 \{w_5, w_6\} < w_0 \{w_4, w_7\} < w_0 \{w_1, w_2, w_3\}$

Figure 4.1: Worlds for a Just So story about Sophie. The proposition $p$ represents “Sophie went to the parade”, $q$ is “Sophie saw Pedro” and $r$ is “Sophie got stuck behind someone tall”. The actual world is $w_0$, and we assume the agent knows this; thus only the similarity relation for $w_0$ (shown below the table) will be needed.

$w_6$ is; equisimilarity is simply the most natural way to fulfil this condition (and will be implied by the causal semantics given in Section 3).

We need to give the agent’s attention ordering $\preceq$ (used to specify which worlds ‘spring to mind’ depending on which atomic formulae she attends to). Here is the ordering we need:

$$w_0 \preceq \{w_2, w_4, w_6\} \preceq \{w_1, w_3, w_5, w_7\}$$

It encodes a default assumption that $r$ is false (that Sophie is not stuck behind anyone tall). This Just So element, too, must be justified at a later date.

Knowing $\preceq$ we can generate $A$ (the worlds entertained) from $\Xi$ (the propositions attended to). If the agent attends only to $\{p, q\}$, her assumption set $A$ is $\{w_0, w_2, w_4, w_6\}$ (that is, she assumes that $r$ is false). This is the sort of assumption-by-omission that we expect when agents fail to attend to specific possibilities. Then the worlds in her cognitive state can be pictured as:

$$w_0 < w_0 < w_0 w_4 < w_0 w_2.$$ Assuming that $w_0$ is the only world in $B$, this clearly supports $B(p \rightarrow q)$. Call this state $\sigma_1$.

If the agent becomes aware of $r$, however, her state expands to include all eight worlds:

$$w_0 < w_0 \{w_5, w_6\} < w_0 \{w_4, w_7\} < w_0 \{w_1, w_2, w_3\}$$

Call this state $\sigma_2$. Now $\sigma_2$ no longer supports $B(p \rightarrow q)$, although it does support $B((p \land r) \rightarrow \neg q)$. And what could trigger such an expansion? The simplest possibility is, hearing “$p \land r \rightarrow \neg q$”. One feature of forward Sobel
sequences is how naturally they read as an argument between two participants, instead of a pair of utterances coming from one agent.

Let $\sigma_0$ be a state of no awareness, where the agent entertains only $w_0$. Then the schematic picture of the possible dynamics looks as in Figure 4.2.

$$
\models X(p \Box q) \\
\n\not\models X((p \land r) \Box \neg q)
$$

Figure 4.2: Awareness states and possible updates. State $\sigma_0$ is the state of no awareness; recall that $X$ is the explicit belief operator. In all states only $w_0$ is in the belief set $B$. The belief updates are all tests, so the states only change because of the awareness component: at $\sigma_0$ and $\sigma_1$ the agent does not attend to $r$, while at $\sigma_2$ she does. The state labelled “$\bot$” has an empty (absurd; inconsistent) belief set.

Note that the shifting strict semantics would produce the same transitions. That is, with the attention dynamics as I have specified, we get exactly the same predictions as from the shifting strict semantics. However, since this is only one possible parameter setting, the present account captures a range of phenomena that the shifting strict semantics cannot.

2.1 · Dinosaurs and tautologies

We capture the counterexamples I gave to the shifting strict semantics (in Section 1.2.1) without major difficulties, on natural assumptions about the orderings involved. In fact, nothing extra is needed to represent the first two examples. For (4-d) I have to do a little more work: I have to convince you that

---

9Unless something truly pathological happens when the agent becomes aware of the possibility that Sophie gets upset; I ignore the possibility.
it is distinctly unlikely that Sophie would get eaten by a dinosaur if she went to the parade. Recall that to get the standard Sobel sequences right, I have to assume that \( w_5 \) and \( w_6 \) are equisimilar to the actual world. This is defensible, on a suitable notion of ‘similarity’ (see Section 3), however the same cannot be said for the corresponding worlds (with and without dinosaurs at the parade) of (4-d)! The same holds (in less spectacular fashion) for (4-c): Sophie being shorter than she actually is need not be taken as seriously as her being exactly as tall as she actually is (again see Section 3).

2.2 · Influence of factual information

(8) a. A: If I had taken the fight and Big Joe was drunk, I might have won.
   b. B: Big Joe wasn’t drunk.
   c. A: Ah. Then if I had taken the fight I would have lost.

It is perhaps not surprising that the system makes the correct prediction for this example, but I mention it in particular as an argument for modular system design. Factual assertion (“Big Joe wasn’t drunk”) removes worlds from the belief set, while the counterfactual test takes the belief set as its starting point when collecting most similar \( \varphi \)-possibilities (“If I had taken the first. . .”). The interplay of these orthogonal systems produces exactly the effect we want, without needing any special rules for the combination.\(^{10}\)

2.3 · Kernel of the account

So far I have given only specific examples of the awareness account in action. Here are the general features. Suppose the Sobel sequence has the form \( p \rightarrow q; p \land r \rightarrow \neg q \). What is required to make this account work?

For Sobel sequences on the Heim pattern (felicitous forward but infelicitous backward) we need three features:

1. Full awareness still justifies \( (p \land \neg r) \rightarrow q \);
2. Full awareness justifies \( p \rightarrow \text{might } r \);
3. Unawareness of \( r \) produces a default assumption that \( \neg r \).

\(^{10}\)At the risk of flogging a dead horse, let me point out that the shifting strict analysis, even if relativised in the obvious way to information states rather than worlds, cannot accommodate this example (as [Fin01] concedes; see his discussion of the “resetting of the modal horizon” by “a rather indirect pragmatic mechanism”). The problem is that the context shift must be assumed to take place for all worlds in the belief set; the modal horizons of \( w \) (where Big Joe is drunk) and \( v \) (where he is not) may gain different worlds, but they both must end up containing counterfactual possibilities where he is drunk. This means that under the shifting strict semantics, factual information can have an effect on counterfactuals . . . but only when that factual information has not already appeared in the antecedent of a previous counterfactual. (Example (8) is predicted unacceptable under the shifting strict analysis, updated to information states, unless you remove “and Big Joe was drunk” from A’s first utterance.)
The first element is non-negotiable: in all Sobel-like examples the complex conditional is, ultimately, acceptable. The second is also in accord with normal intuitions, although advocates of the shifting strict analysis would claim that it is a byproduct of the update, rather than an independent fact revealed by growing awareness. It is the third that drives the account for Heim’s examples, and the combination of the last two that needs the most justification.

This combination throws light on another feature of Heim’s example. There is a temptation to undermine the force of these examples by ‘repairing’ reverse Sobel sequences. The point is most clearly made with respect not to Sophie but to Lewis’s more political example (1-b), expanded to a Heim-style Sobel sequence as follows:

(9)    a. (i) If the US threw all its nuclear weapons into the sea tomorrow, there would be war.
        (ii) But if the US and all the other nuclear powers threw their nuclear weapons into the sea tomorrow, there would be peace.
    b. (i) If the US and all the other nuclear powers threw their nuclear weapons into the sea tomorrow, there would be peace.
        (ii) But if the US threw all its nuclear weapons into the sea tomorrow, there would be war.

Discussing this example Gillies writes in [Gil07, p. 332, footnote 5] that there “may be some temptation” to argue that the sequence should instead be read as in (10).

(10)    a. (i) If only the US threw all its nuclear weapons into the sea tomorrow, there would be war.
        (ii) But if the US and all the other nuclear powers threw their nuclear weapons into the sea tomorrow, there would be peace.

Gillies correctly points out that this pair is no evidence against strengthening the antecedent, and so concludes: “We had better resist such temptations.”

But the temptation itself is interesting data. (9-a-i), when uttered under the assumption that \( \neg r \), expresses exactly the same as (10-a-i). If later conversation makes clear that the utterance was made under such an assumption, then a charitable attempt at paraphrase might naturally make that assumption explicit (even if only as a correction: “I thought you meant that if only . . . ”).

What the first two features above tell us is that in fact these pairs are not (in the Heim cases with infelicitous reversals) evidence against strengthening the antecedent for counterfactuals. Such evidence does exist; (4-d) provides it, for example. But what the Heim observations show is that we need to be very careful to enforce simultaneity if we are looking for such evidence.
In cases where the Sobel pair is genuinely simultaneously true, the second feature of my summary above fails to hold. In these cases (such as the examples in Section 1.2.1 above), the reversed sequence remains felicitous.

This feature gives a constraint on relative similarity of worlds where Sophie does and doesn’t get stuck behind someone tall, and thus indirectly tells us something about the notion of similarity in operation. The ordering semantics is in a sense too flexible: unless we consider only a carefully selected class of possible orderings, we might get stuck with an ordering that considers a world where Sophie at the parade gets stuck behind someone tall less similar to the actual world than one in which she goes to the parade and gets a clear view. Such an ordering would not give us the prediction we want (that the reversed sequence is infelicitous) so we need some way to rule it out. Fortunately, there is a way to pick out the right class of orderings — and even one that has a lot of independent evidence speaking for it. The key is to make use of a causal notion of similarity.

3 · Causal ordering semantics

Causal semantics for counterfactuals emerges from a different challenge to the similarity accounts of Stalnaker and Lewis: the demand that the primitive notion of ‘similarity’ be given some content. The most naïve such notion is probably that $w_1 <_w w_2$ ($w_1$ is more similar to $w$ than $w_2$ is) just in case $w_1$ agrees with $w$ on more proposition letters than $w_2$ does (call this the Hamming distance approach). Clearly this will not work if propositions are directly causally related (in $w$ the trigger was pulled and the gun fired; in $w_1$ and $w_2$ the trigger was not pulled, but $w_1$ where the gun still fired is predicted more similar to $w$ than $w_2$ where it did not). A more subtle problem was pointed out by [Tic76]:

[C]onsider a man — call him Jones — who is possessed of the following dispositions as regards wearing his hat. Bad weather invariably induces him to wear his hat. Fine weather, on the other hand, affects him neither way: on fine days he puts his hat on or leaves it on the peg, completely at random. Suppose, moreover, that actually the weather is bad, so Jones is wearing his hat.

The observation Tichý makes is that “If the weather were fine, Jones would be wearing his hat” seems false in this context, despite (or perhaps because of) the lack of any causal dependency between fine weather and Jones’s habits of hat-wearing.

The basic intuition here is still causal at root: rain causally influences Jones’s hat-wearing, so retracting the rain removes the reason for expecting Jones to wear the hat. Veltman gives a more complicated example:
Suppose that Jones always flips a coin before he opens the curtains to see what the weather is like. Heads means he is going to wear his hat in case the weather is fine, whereas tails means he is not going to wear his hat in that case. Like above, bad weather invariably makes him wear his hat. Now suppose that today heads came up when he flipped the coin, and that it is raining. So again, Jones is wearing his hat.

And again, the question is whether you would accept the sentence ‘If the weather had been fine, Jones would have been wearing his hat.’ This time, your answer will be ‘yes’. [Vel05, p. 164]

Again the intuition is causal, with an additional wrinkle: rain causes Jones to wear a hat, but in the absence of rain the outcome of the coin-flip will also cause him to wear a hat. Veltman models this situation using the notion of the basis of a world: a partial propositional valuation minimal such that it picks that world out against the background of live alternatives. If the set of live alternatives does not include every propositional valuation, then some worlds may be picked out without having to give their ‘full names’ (the complete valuations that identify them). The minimal valuations identifying a world are the bases (there may be several) of that world. In Veltman’s scenario above, the actual world is identified by the basis “raining, heads” (against a background of live alternatives constrained by Jones’s habits of hat-wearing). Skating briskly over several complications, Veltman’s semantics for a counterfactual $\varphi \rightarrow \psi$ are approximated by collecting the worlds satisfying $\varphi$ whose bases differ least (by Hamming distance) from that of the actual world, and testing $\psi$ on that set.

Veltman’s system is not truly causal, although it comes very close. Katrin Schulz has improved on it in this respect, in her dissertation [Sch07] and in ongoing work. We need not be too concerned with the details, especially as her proposal is still being refined (for the latest version at the time of writing, see [Sch]). The core idea is to replace Veltman’s bases with an explicitly causal notion. The causal basis of a world $w$ is again a minimal propositional valuation (a situation), but one that generates the full valuation at $w$ by the action of default causal rules (such as “rain causes Jones to wear a hat”). In Veltman’s system such rules are strict, and act to restrict the universe of possibilities; in Schulz’s formulation the world where it is raining and Jones does not wear his hat simply has a larger causal basis, recording the fact that some causal expectations have been violated.

Schulz derives an ordering from the relative sizes of causal bases. If $b_0, b_1, b_2$ are respectively bases for $w_0, w_1, w_2$, then $w_1 \leq_{w_0} w_2$ iff $b_1 \setminus b_0 \subseteq b_2 \setminus b_0$. That is, if the causal description of $w_2$ diverges from that of $w_0$ more than the causal description $w_1$ does, then $w_1$ is causally more similar to $w_0$ than $w_2$ is. Such divergences may be brought about by differences in the facts (that it is not
raining, while in the actual world it is), or by violations of causal expectations (‘counting miracles’; for example, that Jones does not wear his hat despite the rain, while in more normal worlds the rain causes him to wear it).

3.1 · Reasons
What a causal similarity account brings into clear focus is the importance of reasons for the similarity ordering. Our default expectation is that people at parades see the people parading; to not do so without any reason is a violation of a causal expectation and would count against the world in question in Schulz’s ordering. On the other hand being eaten by a dinosaur would cause Sophie not to see anything; a violation of this expectation would also count against the world.

Where things get interesting is, of course, with the worlds where Sophie gets stuck behind someone tall. This also provides a causal reason for her not to see Pedro, but unlike with the dinosaur we want that reason to be ‘taken seriously’ in the ordering. Schulz’s system allows this, as follows. In the actual world (where Sophie is not at the parade) she is not stuck behind anyone tall. But the reason she is not stuck behind anyone tall is that she is not at the parade (or in a crowd with tall people). If we counterfactually remove that reason (by moving her to the parade) its causal consequences no longer count for similarity of worlds. Technically, the causal basis of the actual world does not include the fact that Sophie is not stuck behind someone tall (because this is a causal consequence of some other fact in the basis, namely that she is not in a crowd); thus different valuations for this fact do not count to differentiate worlds in the ordering.

(Of course the same does not hold for the dinosaur. The reason Sophie is not eaten by a dinosaur (that they are extinct) applies in both the actual world and worlds where she goes to the parade; changing this fact thus counts as a causal expectation violation and differentiates worlds in the ordering.)

It is very natural in ‘counterfactual negotiation’ to explicitly bring up reasons (“Sophie’s really pushy, she never lets anyone block her view”). I will have more to say about this in Section 4.2, but it should be clear enough that a causal account will behave nicely for such examples. If Sophie’s pushiness is causally responsible for her view not being blocked, then a world in which she is blocked violates causal expectations while one in which she is not blocked does not; if she is not pushy (the default assumption), no such violation occurs and the worlds are on an equal footing.

3.2 · A note of hesitation
I have resisted giving a full account of Schulz’s semantics not just because it is still undergoing revision, but also because I suspect that the fit with awareness is not as tight as I would wish. Her formal implementation relies on
constructing a network explicitly representing the causal influences between atomic propositions. The implementation I have sketched above takes the similarity ordering that such a network produces, and filters that through the assumptions of the agent. While technically quite successful, this approach is conceptually a little suspect. The causal network must be very large and intricate, containing many atomic propositions that the agent does not attend to (it must contain the dinosaur, for instance). It would seem more in the spirit of awareness to filter the network itself, rather than the ordering it produces, through the agent’s awareness state. There are however rather formidable technical difficulties standing in the way of this approach; mainly the connection between (possibly complex) assumptions and causal expectations under unawareness is completely obscure, so long as the former are modelled with sets of worlds while the latter are modelled with (something like) formulae.

The notion of closed world reasoning (to which anyway Schulz’s technical implementation is closely related) might provide a strategy to overcome these difficulties. Very many of the causal relations we need for examples like Sophie at the parade have the form of a closed-world rule: If you go to a parade then unless something unexpected happens, you see the people parading. (A dinosaur is unexpected at a parade; if it eats you, unless something unexpected happens you won’t see anything.) The clause “unless something unexpected happens” is at the heart of the closed-world approach: if we are not told that anything unexpected happened, then assume that it didn’t. For causal rules of this form, the relevance of unawareness is clear: the agent need not be aware of all possible exceptional cases in order to reason using the rule.

I have not been able to pursue this hunch further, but the question of the order of explanation seems particularly interesting. It doesn’t take much imagination to see a default rule of this kind not as underpinning or generating counterfactual beliefs but rather as an expression or consequence of them. The interaction with assumption becomes particularly interesting; it is certainly no coincidence that the assumption that Sophie is not stuck behind anybody tall has a structure so similar to the formulation as a closed-world rule.

Unfortunately I must leave such speculations, to return to the proposal I am certain enough of to wish to defend. I have claimed as a weakness of the shifting strict semantics that it applies only to counterfactuals (or, on the most charitable reading, only to conditionals). In the next section I will show some non-conditional examples that appear to share features with the Sobel sequence data above, and to which the awareness account can be applied.

4 · Beyond counterfactuals
Arguing directly against the shifting strict analysis, Moss [Mos07] gives a number of examples of Sobel-like patterns that do not involve counterfactuals.
Indeed, part of my argument in this chapter is that it may have been a mistake to term the pattern Heim noticed a ‘Sobel sequence’. Sobel’s observation was that counterfactuals do not support strengthening of the antecedent; if we accept the ordering semantics, then that fact has nothing to do with awareness. Heim’s observation, I would say, is that truth-value judgements of counterfactuals can change under conditions of changing awareness; that fact has very little inherently to do with counterfactuals!

So I need to show that my account can be extended to the non-counterfactual cases without difficulties. What are these cases?

Perhaps most famously there is [Lew79, pp. 354–355]:

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

Here, at least, I need do no extra work: modals of possibility and necessity must be sensitive to what options are being attended to, as I have already argued for might in the previous chapter. But there are a number of non-modal examples which are a little trickier.

[Wilo8] proposes a shifting strict analysis for indicative (rather than counterfactual) conditionals, for cases like the following (I adapt the example a little, innocently I hope):

\[(11)\]
\[\text{a. (i) If Oswald didn’t kill Kennedy that day in Dallas, somebody else did.} \]
\[\text{ (ii) But if the \text{kgb} kidnapped Kennedy and his death was faked, nobody killed him.} \]
\[\text{b. (i) If the \text{kgb} kidnapped Kennedy and his death was faked, nobody killed him that day in Dallas.} \]
\[\text{ (ii) ?But if Oswald didn’t kill Kennedy, somebody else did.} \]

Moss points out the analogy with a famous example: \(^{11}\)

\[^{11}\text{Moss admits that “our familiarity with zebra examples can create unwanted noise in our judgements about them.” She suggests as a fresh alternative the exchange given in (6).} \]
Chapter 4 · Case study: Sobel sequences

(12)  a. (i) That animal [a zebra] was born with stripes.
      (ii) But cleverly disguised mules [with stripes painted on] are not
           born with stripes.

     b. (i) Cleverly disguised mules are not born with stripes.
           (ii) ?But that animal was born with stripes.

The original point of these examples was that the possibility being introduced apparently undercuts the ability of the speaker in (12-b-ii) to know what is here given as his utterance. That same property seems to make the utterance infelicitous, at least if it is imagined as not stressed in any way (the stressed example belongs with (13) below).

A devotee of update semantics should feel a little uncomfortable with these examples. On the face of it, (11-b-ii) and (12-b-ii) are nothing but bald assertions; the hearer’s state should be updated with their information content, and this will cause no problems unless the hearer holds an explicitly contradictory belief. An utterance that would take the hearer to the absurd state is pragmatically ruled out, but that is certainly not the problem here... so what is going on?

4.1 · Speaker expertise

The answer is that we are also used to the speaker knowing what they are talking about. The notion of ‘an update with the information content of an utterance ϕ’ assumes that ϕ contains information: that what it says is true, or describes the world truly. The very construction of these examples leads us to believe that the utterer of (11-b-ii) or (12-b-ii) is not in a position to know that her utterance is true.

Moss describes this as a norm of assertion: speakers should not make assertions that are incompatible with any salient possibilities that they are not in a position to rule out. She also points out (footnote 10, pg. 11) that “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.”

Awareness-relative epistemology will have to remain a project for future work, however some bounds can be fairly confidently assumed. While I hesitate to say how damaging unawareness may be to an agent’s actual knowledge, she certainly does not believe that she knows that ϕ (say) if she entertains and holds possible some contingency incompatible with ϕ. And while I hesitate to say whether knowledge should be a norm of assertion, I’m much more confident that believing to know should be.\(^{12}\) The awareness-relative reading of all these

\(^{12}\)All this hesitation is tiresome. But the unawareness perspective seems to tend inexorably towards a particular kind of relativism: According to A, B knows that ϕ. The observer/judge A is needed to set the standards of awareness; otherwise B either knows too much (if her own standards are also normative; “the epistemic efficacy of stupidity”, as Catherine Elgin puts it [Elg88]) or too little
examples is that the second speaker of the reverse sequence does not believe she knows what she is asserting, which should explain the infelicity.

This effect is heightened, I think, by a subtle property of the linguistic presentation (as forward and reverse pairs of example sentences). That is the temptation to read (11-b-ii) as identical to (11-a-i), and (12-b-ii) similarly as identical to (12-a-i). By this I mean not only that they have the same truth conditions, but that we seem to interpret them as if they were prompted by the same epistemic state on the part of the speaker. The state most naturally assumed to prompt (12-a-i) (i.e., the assumption that no clever disguises are in effect) is clearly insufficient to justify (12-b-ii). If the examples were not presented back-to-back, though, we would not be tempted to assume that the epistemic state of the speaker was the same in both cases: if he feels licensed to assert (12-b-ii) then it is because he is sure that there are no clever disguises involved. The examples are of course carefully picked to make this certainty insufficiently justified for a claim of (‘strong’, ‘philosophical’) knowledge, but knowledge is not our primary concern here. Compare these variations (essentially read as adversarial conversations, but on that reading completely felicitous, as far as I can tell):

\[
\begin{align*}
(13) & \quad \text{a. (i) If the \textit{kgb} kidnapped Kennedy and his death was faked,} \\
& \quad \text{ nobody killed him that day in Dallas.} \\
& \quad \text{(ii) I’ve looked through the \textit{kgb} historical archives. If Oswald} \\
& \quad \text{ didn’t kill Kennedy, somebody else did.} \\
& \quad \text{b. (i) Cleverly disguised mules are not born with stripes.} \\
& \quad \text{(ii) But if you look closely, I’m quite certain that you’ll agree with} \\
& \quad \text{ me: that animal was born with stripes.} \\
& \quad \text{c. (i) Cars get stolen all the time here in New York City.} \\
& \quad \text{(ii) But \textit{my} car is around the corner. I’m naturally lucky.}
\end{align*}
\]

I include this last example particularly to undercut the idea that a knowledge norm of assertion has any direct relevance for our judgements of felicity. What seems to be at issue is whether (we are satisfied that) the speaker \textit{believes} she has the required knowledge. If she believes this erroneously (as in (13-c-ii)) we may disagree with her statement, but we feel no temptation to censure her as an uncooperative speaker.

To sum up: I agree with Moss that what is at stake is properly addressing the alternative possibilities that have been made salient, and also that this has something to do with norms of assertion. However we part company when

\(\text{if the normative standards include awareness of ‘everything there is’}.\) I am not epistemologist enough to know what to do with this position, except to wonder where it leaves the notion of knowledge ‘in the abstract’, without an explicit ascriber. Hence the hesitations. Section 2.1 of Chapter 7 contains some further hesitations on the same subject, as a suggestion for possible further work.
it comes to what the connection actually is. As far as I am concerned it is a condition of rational belief formation that you not discount salient alternatives without reason; the relevant norm of assertion is “Assert only what you believe you know”, and the pragmatic infelicity markings above point to a suspicion of irrational belief formation rather than of deception or similar deliberate violation of norms.

At this point we might wonder, can we play the same game with counterfactuals? Can we rehabilitate reverse Sobel sequences just by taking opinionated speakers? Interestingly, the answer seems to be no. The reasons why not have to do with the extra sensitivity of counterfactuals to changes in awareness.

4.2 · Uncertainty about counterfactuals

Suppose one is uncertain whether it is really the case that if Sophie had gone to the parade she would have seen Pedro. It seems there are two kinds of uncertainty one could be suffering from, shown in (14). (14-a) is rooted in lack of knowledge about what is the case, while (14-b) is rooted in lack of knowledge of what would have been the case in the counterfactual scenario.

(14)  a. I don’t know whether Pedro in fact showed up at the parade; so I don’t know whether Sophie would have seen him if she had gone.
    b. Sophie is kinda short, and there were lots of tall people at the parade, she might very well have been stuck way back in the crowd and not seen anything; so I don’t know whether Sophie would have seen Pedro if she had gone.

The first example is standard informational uncertainty; we might think of the second as METAPHYSICAL UNCERTAINTY, because it is imposed on us by the metaphysically indeterminate nature of counterfactual alternatives. A similar notion arises in considering the indeterminacy of the future. We might want to represent my uncertainty about the outcome of a coin flip differently depending on whether the coin has not yet been flipped or whether it has been flipped but I have not yet looked at it. Stalnaker famously does not believe in metaphysical uncertainty in this sense, since it leads to violations of conditional excluded middle (at the metaphysical level of truth-conditions); instead he would represent the coin-flip with epistemic uncertainty between world-histories (incorporating future events).

To some extent the decision is an aesthetic one: what individuating conditions for worlds are the most natural in my setting? This is particularly the case for the temporal sequence case, since many branching-time models can be freely inter-translated with models based on world/time pairs or similar constructions. There might be more to the issue for counterfactuals, though. A causal semantics along the lines of Schulz’s model, for instance, is designed to
Beyond counterfactuals

generate worlds equidistant from the actual world in cases like Tichý’s above: in the absence of reasons for Jones to wear his hat, he may just as well wear it as not wear it. We will see some more examples below where metaphysical uncertainty does some work for us.

Perhaps we can even distinguish between the two notions empirically. To my ears (15-a) is somewhat marked, while (15-b) is fine.

(15) a. ?I don’t know whether Pedro was at the parade, so Sophie wouldn’t necessarily have seen him if she went.
   b. Sophie is kinda short, and there were lots of tall people at the parade, so she wouldn’t necessarily have seen Pedro if she had gone.

Here is the picture suggested by the distinction (simplified by omitting several possible valuations). If \( B \) represents the agent’s state of belief, she would say (14-a), whereas if \( B' \) is her state of belief then she would say (15-b).

![Diagram](image)

**Figure 4.3:** Different kinds of counterfactual uncertainty. (Some worlds are shown twice, to make the ordering relations easier to read.) Agents in either of belief states \( B \) and \( B' \) fail to believe \( p \Diamond q \), however an agent in \( B \) can come to believe \( p \Diamond q \) via an update eliminating worlds from her belief set, while an agent in \( B' \) cannot.
Although neither information state currently supports $p \square \rightarrow q$, there is an important behavioural difference between the two: $B$ can be transformed by an informational update into a state that \textit{does} support the counterfactual, while $B'$ cannot. Suppose for instance that the agent learns that Pedro was at the parade (see Section 2.2). Then $w_0$ is eliminated, and the resulting state supports $p \square \rightarrow q$.

In Figure 4.3 we are to some extent comparing apples with oranges, however: the state rooted at $B'$ assumes that Pedro was at the parade, while the state rooted at $B$ assumes that Sophie would not get stuck behind tall people. If we combine the two into a single state of attention, the result looks somewhat different:

If the agent assumes $\neg r$ then we get the state labelled $B$ in Figure 4.3; if she assumes $s$ then we get the $B'$ of Figure 4.3. But if she is aware of everything, then even knowing that Pedro was at the parade she remains uncertain whether Sophie would have seen him if she had gone.

If we give anything like a standard update semantics to counterfactuals, then they can be used just like standard assertions to resolve \textit{informational} uncertainty (to eliminate worlds from $B$) but they cannot be used to resolve \textit{metaphysical} uncertainty. The test semantics we have been using so far does not even allow informational updates, but it does seem reasonable that a counterfactual can carry information (“If Sophie had gone, she would not have seen Pedro” is a roundabout way to tell an agent in state $B$ above that Pedro was not at the parade, on this account). However this possibility is almost completely irrelevant for the vanilla Sobel examples: the ‘null context’
of linguistic examples, combined with a reason-based ordering semantics, virtually guarantees that the uncertainty involved will be metaphysical rather than informational.

**Definition 4.5**: Ordering semantics for attention models (counterfactuals as informational updates). Let $M = \langle W, \Omega, V, S^W \rangle$ be a model of reality augmented with an ordering frame, and $\sigma = \langle A, B, \Xi \rangle$ a state.

\[
\langle A, B, \Xi \rangle [\varphi \square \Rightarrow \psi]_b^M =_d \langle A, B^+, \Xi \rangle
\]

where

\[
B^+ = \{ w \in B ; \langle A, C_w, \Xi \rangle \models^M \psi \}
\]

and for each $w \in B$,

\[
C_w = \min_{\leq_w} A \upharpoonright \varphi
\]

This is of course nothing but the standard update formulation for assertions: keep the worlds from $B$ that satisfy the formula in question, and throw away the rest.

We have now two update rules for the counterfactual conditional: one a test, and one a substantive (informational) update. One could wonder, do we need them both? I think we do, so long as we make two commitments:

1. counterfactuals can be informative, and
2. the *might* of a *might*-conditional scopes semantically under the conditional operator.

The first principle is uncontroversial; the second is certainly open to debate (an alternative is to stipulate that $\varphi \square \Rightarrow \psi$ be analysed as $\text{might}(\varphi \square \Rightarrow \psi)$). If we uphold these two principles, though, we definitely need two kinds of update. The first principle requires (something like) the informational update given in Definition 4.5, but the combination of this definition with the second principle makes *might*-conditionals too strong. In example (7), for example, my utterance would (wrongly) provide the information that Big Joe is drunk.13

Assume that we do need both definitions. Then how does an agent, hearing a particular utterance of a counterfactual conditional, choose which one to apply? I think the decision is essentially a pragmatic one, driven by considerations of speaker expertise. Is there a point of epistemic uncertainty at stake, on which the speaker can be taken to be expert? If so, we may apply the

---

13Definition 4.5 is completely out for those who believe in conditional excluded middle: it follows from $\varphi \square \Rightarrow \psi$ that $\varphi \square \Rightarrow \psi$, if no metaphysical uncertainty is possible. The only solution seems to be to scope *might* above the counterfactual; whether this is acceptable (quite apart from compositionality principles) depends on whether you believe a *might*-conditional can ever be informative.
Chapter 4 · Case study: Sobel sequences

informational update; if not, we had better take the test. There may be more pragmatic reasoning involved, such as whether the counterfactual utterance is a particularly roundabout way of conveying a simple message — I will give an example in a moment, but first now let us see what happens when we try a reverse Sobel sequence in a context where it is clearly epistemic uncertainty at stake (and thus where the informational update would be expected).

(16) a. A: Perhaps Pedro wasn’t at the parade. If Sophie had gone and Pedro wasn’t there, she wouldn’t have seen him.
   b. ?B: If she had gone she would have seen him.

My intuition is that (16-b) is somewhat marked, but much more acceptable than (3-b). Part of the markedness is probably due to the fact that speaker B appears to want to convey “Pedro was there” but does to in rather a roundabout way. Moss gives an example which avoids this problem, and also involves informational rather than metaphysical uncertainty:

Suppose John and Mary are our mutual friends. John was going to ask Mary to marry him, but chickened out at the last moment. I know Mary much better than you do, and you ask me whether Mary would have said yes if John had proposed. I tell you that I swore to Mary that I would never 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:

(18) 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. [Moss07]

\[14\] [Vel05] also has dual updates (informational and test) for the counterfactual; for him the choice of which to apply, also pragmatic, rests on whether the laws in play are fully known or not. This is roughly comparable to the question of whether the agent is aware of the causal dependencies at stake or not, although it raises the vexed issue of how an agent unaware of some law (or causal dependency) should recognise that unawareness and alter her behaviour accordingly.
\[15\] It is interesting to note that “But” is completely out here. I have no theory to account for this. As far as I know, nobody has systematically explored in which Sobel-like configurations “but” is permitted, obligatory, or prohibited. For what it’s worth, “But” also seems to be out in the dinosaur variant:

(17) a. If Sophie had gone to the New York Mets parade and been eaten by a dinosaur, she would not have seen Pedro Martínez.
   b. (? But) If Sophie had gone to the parade, she would have seen Pedro.
Here the feature of the real world to be communicated is Mary’s attitude to John, as already conveyed to the speaker in the actual world.

This brings up an interesting point. The best description of that attitude is probably a conditional one: “will not marry him if he asks her”. But it is only a very short step to a counterfactual attitude: “would not have married him if he had asked her”. Indeed, this might very well be what Mary has told the speaker: “For a moment yesterday I thought John was going to propose. I would have said no, though.” The feature of the actual world that is being conveyed is a counterfactual disposition. And if Mary may have this counterfactual disposition to reject proposals, why cannot Sophie have a counterfactual disposition to avoid tall people?

In fact I think she can; the two that spring to mind are “Sophie is really pushy and doesn’t let anyone get in the way of her view” and “Sophie is really short and always gets a bad view at parades”. Either of these, if taken as potential facts in the actual world, would suffice to turn the metaphysical uncertainty about which of $w_3$ and $w_4$ takes priority into informational uncertainty (about whether the disposition holds in the actual world). There are two contrasts with the proposal example, however.

The first contrast has already been discussed: Moss’s example deliberately makes the speaker an expert on the disposition under discussion, whereas a typical Sobel example undermines the speaker’s potential for expertise.

The second contrast is more interesting. This is that both of the dispositions Sophie might hold are ‘marked’: they need particular names (“pushy”, “unusually short”) and the most natural awareness model provides global assumptions (“not pushy”, “not unusually short”) for both of them. In the null context of a Sobel example, it seems we naturally assume (and let our agents also assume) that these dispositions do not obtain. In contrast, the proposal example is explicitly about Mary’s disposition to accept John’s proposal; even if we give the disposition a name that is not being explicitly attended to yet (“love”?) there is no temptation whatsoever to hold assumptions about the valuation that name should get.

Our model is almost capable of using this second contrast in order to fully represent the difference between these two situations. The point of leverage is the notion of belief-attention-consistency. The picture corresponding to Figure 4.4 but for Mary’s proposal is given overleaf, in Figure 4.5.

Now suppose that the agent is unaware of $x$, but entertains both $w_1$ and $w_2$. By the definition of belief-attention consistency, she may not eliminate either $w_0$ or $w_1$ from $B$! Distinguishing between $w_0$ and $w_1$ would violate the clause saying that substantive beliefs may not separate worlds that ‘look the same’ through the lens of attention: those that make the same propositions of $\Xi$ true. This is a somewhat underhand trick, as can be seen if we recall the reason for
the definition: the agent should be able to describe the difference that underlies her substantive belief. In this case she could do so, even given her limited awareness: the formula “$p \Box \neg r$” will do the trick quite nicely. However it is more reasonable if we require the agent to also justify her beliefs: the only justification she could give would have to refer somewhere to $x$.

It is now only a short step (albeit one we will not take formally) to distinguishing the two cases. In trying to understand what could justify the speaker in believing “$p \Box \neg r$”, our agent spontaneously becomes aware of $x$: the notion of planning to accept a proposal is so similar to that of accepting a proposal that this is entirely natural. Now she may form the belief without her state thereby being belief-attention inconsistent. In contrast, in trying to understand why the speaker believes that Sophie would not have been stuck behind anyone tall, our agent fails to imagine anything that could justify the belief; she concludes that something is wrong (either the speaker is making unjustified claims, or she is unaware of something she should be aware of), she rejects the update and instead asks for more details (“What makes you think that?”).

Unfortunately this last must remain a Just So story, since we have no mechanism for the spontaneous association of ideas.