Proceedings of SALT 25: 1–17, 2015

Informative counterfactuals[*]

Adam Bjorndahl
Carnegie Mellon University

Todd Snider
Cornell University

Abstract A single counterfactual conditional can have a multitude of interpretations that differ, intuitively, in the connection between antecedent and consequent. Using *structural equation models* (SEMs) to represent event dependencies, we illustrate various types of explanation compatible with a given counterfactual. We then formalize in the SEM framework the notion of an *acceptable explanation*, identifying the class of event dependencies compatible with a given counterfactual. Finally, by incorporating SEMs into possible worlds, we provide an update semantics with the enriched structure necessary for the evaluation of counterfactual conditionals.

Keywords: Counterfactuals, structural equation models, informativity, causal dependency, update

1 Introduction

Counterfactuals are used to talk about things we know to be false, as well as things we are simply unsure of. For example, (1) can be used if the speaker knows that Alice did not attend the party, or if the speaker is unsure whether she did; (2) would typically be uttered by a speaker who believes that the movie was not, in fact, any good.

(1) If Alice had gone to the party, Bob would have stayed home.

(2) If the movie had been any good, I wouldn't have fallen asleep.

(3) Even if there hadn't been traffic, Francis still would have been late.

Nonetheless, despite describing states of affairs that are counter to fact or uncertain, counterfactuals are used to communicate about the actual world. This happens in two ways. First, counterfactuals can encode information about the truth values of events in the actual world. For instance, (1) often implies that Alice in fact did not attend the party, and that Bill did. The *Even...still* construction in (3) communicates

[*] Our thanks to Christina Bjorndahl, Sarah Murray, Will Starr, the Cornell Semantics Group, audiences at PHLINC2, LGM, NASSLLI 2014, and SALT 25, and many anonymous reviewers for insightful discussions, questions, and critiques. All errors are our own. Adam Bjorndahl's work was supported in part by NSF grant CCF-1214844 and by ARO grant W911NF-14-1-0017.

that in fact there *was* traffic, and in fact Francis *was* late. Accommodation of this sort of presupposed content (as in Stalnaker 1974) is one way in which counterfactual statements can serve to update our knowledge of the actual world.

In this paper, however, we focus on a second sense in which counterfactual conditionals are informative: namely, they encode information about a connection between the antecedent and consequent.

1.1 Acceptable explanations

Generally speaking, the felicitous utterance of a counterfactual conditional is compatible with a multitude of possible connections between the antecedent and the consequent. Consider, for example, the utterance in (1); while it certainly encodes some sort of connection between Alice's being at the party and Bob's being at the party, it is silent about the specifics. Perhaps Bob is avoiding Alice, as in (4).

(4) X: If Alice had gone to the party, Bob would have stayed home.
Y: Why?
X: He owes her some money.

Of course, there are many explanations for why Bob might be avoiding Alice, of which (4) represents only one: perhaps Bob dislikes Alice, or he is just shy; perhaps Alice is Bob's committee chair, to whom he owes a draft. But the space of possible explanations is even wider than this: consider the discourse in (5).

(5) X: If Alice had gone to the party, Bob would have stayed home.
Y: Why?
X: They hate these sorts of functions, so they take turns going.

In this case, Bob isn't avoiding Alice at all; rather, there is some shared reason for their attendance choices. (5) explicitly provides one such reason, but again there are many others. Perhaps they share a budget and can't both afford to go. Or perhaps they share a child and must ensure that he is supervised.

The multitude of explanations compatible with an utterance of (1) seems endless. Yet it also seems to be governed by certain rules, as shown by the infelicitous exchanges in (6).

(6) X: If Alice had gone to the party, Bob would have stayed home.
Y: Why?
X_1: # Bob sometimes doesn't attend these parties.
X_2: # Alice is dead.
X_3: # She owes him some money.[1]

[1] Of course, with a few contortions one can read this response as felicitous (e.g., perhaps Bob does

In this paper, we identify the class of *acceptable explanations* for a given counterfactual by formalizing the connection between antecedent and consequent. As we have seen, the truth value of a counterfactual alone is not enough to determine the nature of this connection. Moreover, the standard semantic accounts of counterfactuals using *similarity relations* gloss over this issue altogether by abstracting away from the mechanisms that connect the truth of the antecedent to the truth of consequent. We present a model for counterfactuals that embraces their explanatory underspecification by explicitly modeling dependencies between events, and we explore some of the insights that arise from this way of thinking. In particular, our definitions in §3.2 imply that the three infelicitous responses in (6) exhaust the semantic obstructions to an explanation being acceptable.

2 Overview

2.1 Informativity

Traditionally, semanticists and philosophers of language call something *informative* when it meaningfully reduces the *context set*—the set of live possibilities as to how the world might be, with respect to the current discourse. Such possibilities are often conceptualized as *possible worlds*, in which case the context set might be thought of as the set of worlds that could plausibly be the actual world. Roughly speaking, then, an utterance is *informative* if it eliminates some (but not all) candidate worlds under consideration. A simple utterance of a declarative sentence which encodes the proposition φ, for example, partitions the universe of possible worlds into φ-worlds and $\neg\varphi$-worlds, and then rules out all the $\neg\varphi$-worlds, effectively removing any $\neg\varphi$-worlds still in the context set.

We subscribe to essentially the same notion of informativity, but because we aim to explicitly model the connections between events, our analysis relies on a richer notion of possible worlds than is standard. Specifically, in addition to encoding the *truth values* of events, our worlds will encode *the dependencies among them*, following Starr (2012) and Briggs (2012). This enrichment provides an additional means by which to discriminate among possible worlds. As is the case classically, we can form partitions based on the truth value of a specific event, but now we can also partition based on the existence (or non-existence) of a particular dependency. That is, if an utterance asserts the existence of some dependency d, we can model that assertion's update as dividing the universe into d-worlds (worlds with that particular dependency) and $\neg d$-worlds (worlds without that particular dependency), and then remove worlds in the standard way.

not *want* Alice to settle her debt), but this returns us to a scenario in which Bob is avoiding Alice, as opposed to the natural interpretation in which Alice is avoiding Bob.

In sum, we rely upon the standard notion of informativity, but with an additional dimension along which to partition. A dependency-conveying assertion is informative if and only if updating by that dependency meaningfully reduces the context set.

2.2 The similarity approach

The standard approach to counterfactuals, often called the *similarity* approach, relies on a *similarity relation* to encode a notion of closeness between worlds (Lewis 1973, 1979). Roughly speaking, a counterfactual conditional is true on this account just in case the closest worlds where the antecedent is true are worlds where the consequent is true; however, nothing explains *why* this is the case—no explicit mechanism connects the two propositions.

Our analysis relies on no such relation: all of the work done by the similarity relation is accomplished instead by reference to the dependencies among events encoded within individual worlds. Moreover, this dependency structure permits a finer-grained analysis of the truth conditions for counterfactual conditionals than is possible using similarity relations, allowing us to explore the rules governing acceptable explanations.

2.3 Structural equation models

We make use of *structural equation models* (SEMs) to represent event dependencies. This framework goes back as far as Wright 1921, but we work with the version formalized by Pearl (2000). An SEM can be pictured as a graph consisting of nodes, edges, and labels. Each node, represented as a lettered circle, stands for an event variable. Each edge, represented as an arrow between two nodes, reflects a directed dependency between events. Finally, the labels encode the specific nature of these dependencies. In this paper we restrict our attention to SEMs that are two-valued (i.e., variables are either true or false) and deterministic (i.e., the value of a child node is a *function* of the values of its parents, as opposed to a relation specified probabilistically).

In §3.2 we provide formal definitions; for now, we focus on examples. Consider first the SEM depicted in Figure 1, which includes two abstract events, X and Y, and encodes a dependency of Y on X. More precisely, the label stipulates that Y gets the same truth value as X: if X is true then Y is true, and if X is false then Y is false. This is in keeping with a general principle of (deterministic) SEMs, namely that the value of a child node is determined entirely by the value(s) of its parent node(s).

Consider now the SEM depicted in Figure 2, and recall the counterfactual conditional in (1).

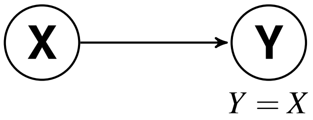

Figure 1: A simple structural equation model.

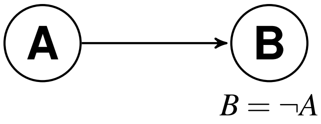

Figure 2: The *direct cause* explanation of (1).

(1) If Alice had gone to the party, Bob would have stayed home.

If we let the variable A stand for "Alice goes to the party", while B stands for "Bob goes to the party" (and we assume for simplicity that Bob not going to the party is equivalent to Bob staying home), then this SEM represents one particularly simple way of spelling out the relationship between Alice and Bob's party attendance: Alice's attending the party *directly causes* Bob to stay home.

In some contexts, this level of detail may be appropriate. But in other cases, such an explanation will fall short of the conversationally appropriate standards for specificity (we discuss this further in §4.1). What is it about Alice attending the party that results in Bob's absence? Insofar as such details are relevant to the discourse, they ought to be encoded in the common ground. In this regard, SEMs furnish the needed structure: more elaborate explanations can be captured by more sophisticated SEMs; moreover, different types of explanations can be characterized by the structural properties of the SEMs that realize them.

2.3.1 Additional causes

One plausible explanation for (1) is that Bob hates Alice and does not want to spend time with her; because of this, he avoids parties that she attends. In keeping with this intuition, we can represent Bob's hatred-driven-avoidance of Alice as an extra node H in the model, as in Figure 3.

In this model, the consequent (Bob's attendance) is still dependent on the antecedent (Alice's attendance), but not solely so. As we can see from the equation that describes how B inherits its value, the truth value of the consequent now depends

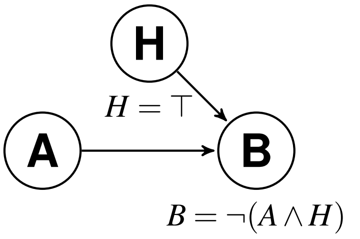

Figure 3: An *additional cause* explanation of (1).

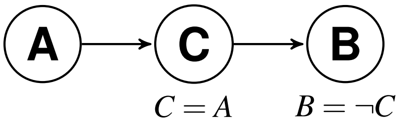

Figure 4: An *intermediate cause* explanation of (1).

both on the antecedent and on an *additional cause*. Specifically, Alice's attendance causes Bob to stay home *under the background assumption that Bob hates Alice*. Note the importance here of fixing Bob's hatred (represented by the structural equation $H = \top$, where $\top$ is the symbol for the logical constant that is always `true`) in order to be able to infer his absence from Alice's attendance.

Of course, there is nothing special about the additional cause in this example being Bob's hatred—one can easily imagine a variety of background conditions under which Alice's attendance results in Bob's absence. In this way, the SEM in Figure 3 can be viewed as a template for a general *type* of explanation, namely, those in which certain background events act as preconditions for the required antecedent-consequent dependence.

2.3.2 Intermediate causes

Another type of explanation for (1) works by mediating the dependence of B on A by some third event. For example, suppose that Alice brings her cat with her wherever she goes, and that Bob is deathly allergic to cats and avoids them at all costs. In this case, Alice's attendance at the party still leads to Bob's absence, but only because of the mediating factor of her cat. This particular scenario is captured by the SEM depicted in Figure 4.

Here, the node C represents the cat being at the party, while the structural

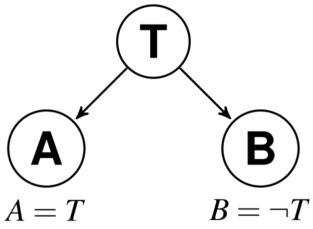

Figure 5: A *common cause* explanation of (1).

equations $C = A$ and $B = \neg C$ encode the fact that the cat travels with Alice and that Bob avoids the cat, respectively. Once again, there is nothing special about the intermediate cause being a cat; this SEM is simply one of many instantiations of a general type of explanation characterized by the presence of intermediaries between antecedent and consequent.

2.3.3 Common causes

There is a third type of explanation importantly different from those considered so far, in that the consequent need not depend on the antecedent at all. Rather, both the antecedent and the consequent depend on some common cause that determines (or at least influences) both of their values. For example, returning to Alice and Bob, we might imagine that they flip a coin to decide who attends the party. This is captured in Figure 5.

Here, the node T represents the event that the coin comes up tails; the labels then guarantee that Alice attends the party if and only if the coin comes up tails, and Bob attends the party if and only if the coin comes up heads. Note that although this scenario involves no causal path from A to B, it is nonetheless intuitively compatible with an utterance of (1).

2.3.4 More complicated models

Of course, several types of explanations may operate simultaneously in a given SEM. Figure 6 provides one example of a more elaborate causal structure that might nonetheless count as an explanation for (1). Indeed, it is easy enough to come up with a plausible story for such a model: a coin is flipped (T) to determine whether Alice (A) or Ursula (U) will attend the party; meanwhile, Vivian (V) avoids Ursula, and Bob (B) avoids Vivian provided he has not watched the latest episode of his

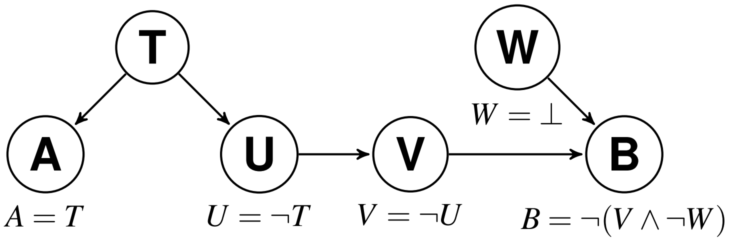

Figure 6: A more elaborate explanation of (1)

favorite show (W), since Vivian follows that show as well and Bob doesn't want to hear any spoilers. A given counterfactual may be compatible with any number of explanations of varying degrees of complexity; exactly which SEMs count as *acceptable explanations* is the subject matter of §3.2 (see also §4.1 for a discussion of other limiting factors).

2.3.5 Backtrackers

Finally, it is instructive to consider what might appear to be another way to ensure the right kind of covariance between antecedent and consequent: rather than having the consequent depend on the antecedent, we can reverse this relationship and model the antecedent as depending on the consequent. Consider, for instance, the SEM depicted in Figure 7. This sort of model is what has been referred to in the classical philosophical literature as a *backtracker* (Lewis 1979): the consequent is the cause while the antecedent is the effect.

We have accepted various explanations for (1) in which Alice's party attendance, one way or another, causes Bob's absence. But explanations that reverse this dependency do not seem to be viable: in any context where it is understood that Alice is avoiding Bob, for example, (1) is infelicitous. Some care must be taken, therefore, in identifying the space of *acceptable explanations*; that is, those explanations compatible with a given counterfactual conditional. It is certainly too much to insist that there exist a causal path from antecedent to consequent—such a condition would rule out the common cause type of explanations discussed above. Instead, we propose the somewhat weaker requirement that there exist *no* causal path from consequent to antecedent; this proposal is formalized in §3.2.

It is interesting to note, however, that the backtracking explanation does become available (and even preferred) under the right syntactic conditions, namely the double-auxiliary construction:

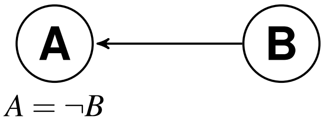

Figure 7: A classical backtracking explanation.

(7) If Alice had gone to the party, Bob would have had to have stayed home.[2]

Understanding counterfactuals via SEMs provides insight into this linguistic feature by identifying a natural class of semantic objects—SEMs like the one in Figure 7—that correspond to the syntactic construction in (7). Moreover, our exploration of the different types of explanation allows us to distinguish two notions that seem to have been conflated in the literature. As noted above, in many classical philosophical accounts a backtracker describes a counterfactual conditional in which the consequent causally or temporally precedes the antecedent. By contrast, in more recent psychological literature, the term is applied to any counterfactual that relies on *upstream reasoning*, that is, that invokes an explanation involving something causally or temporally prior to the antecedent (Edwards & Rips 2012; Rips & Edwards 2013), as for example in Figure 5. These two notions of backtracking behave quite differently, as evidenced by the fact that the latter, but not the former, yields an acceptable interpretation of (1). To our knowledge, this distinction has not been spelled out before now.

3 Formalism

3.1 Underspecification, not ambiguity

We begin by briefly arguing that the multiplicity of interpretations for a given counterfactual is best analyzed not as *ambiguity*, but rather as *semantic underspecification*. Ambiguity is identifiable through the standard VP ellipsis test (see Asher, Hardt & Busquets 2001, and the many references therein): genuine ambiguity, as in words like *bank* (river bank vs. financial institution), is impossible under VP ellipsis. In the sentences in (8), for example, Janine and Kevin must both be at the same sort of bank.

[2] Or, said differently, "For Alice to have gone to the party, it would have to have been the case that Bob stayed home."

(8) a. Janine went to the bank, and so did Kevin.

b. Janine went to the bank, and Kevin did, too.

We can apply the same test to counterfactuals, as in (9).

(9) a. If Alice had gone to the party, Bob would have stayed home, and so would Eve (have).

b. If Alice had gone to the party, Bob would have stayed home, and Eve would've, too.

Note that in both versions of (9), Bob and Eve can have very different reasons for not attending the party, even reasons that cut across distinct types of explanation. For instance, we might have a common cause type explanation for Bob (e.g., a coin toss), and an intermediate cause type explanation for Eve.[3] As such, we formalize our analysis of counterfactuals using a single, underspecified semantics, rather than appealing to ambiguity.

3.2 Structural equation models

Here we provide a rigorous development of the SEM framework we employ, and use it to formalize our proposed semantics for counterfactuals. A **structural equation model** is a triple

$$\mathscr{M} = (\mathsf{Var}, \mathsf{End}, \{\varphi_X : X \in \mathsf{End}\})$$

where

- Var is a finite set of *variables*;
- $\mathsf{End} \subseteq \mathsf{Var}$ is the subset of *endogenous* variables;
- for each $X \in \mathsf{End}$, φ_X is a Boolean expression over Var such that $X \notin dom(\varphi_X)$.

Here, a *Boolean expression over* Var is any expression built from the variables in Var by closing under the Boolean connectives in the standard way; more precisely, it is any expression generated by the grammar

$$\varphi ::= X \mid \top \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi \rightarrow \psi,$$

where $X \in \mathsf{Var}$. We write $dom(\varphi)$ to denote the set of all variables occurring in the expression φ; the $\mathscr{M}$**-parents** of $X \in \mathsf{End}$ are precisely those variables in $dom(\varphi_X)$.

[3] Notably, though, it seems quite difficult to get a reading of (9) where Bob is avoiding Alice but Alice is avoiding Eve—that is, mixing a classical backtracker, as in Figure 7, with a non-backtracking interpretation. That is to say, the classical backtracking interpretation is a different *reading*.

The $\mathscr{M}$**-ancestor** relation is the transitive closure of the $\mathscr{M}$-parent relation, and we write $Y \prec_{\mathscr{M}} X$ to denote that Y is an $\mathscr{M}$-ancestor of X. We omit $\mathscr{M}$ as a prefix and subscript when it is safe to do so. Note that the presence of the constant formula $\top$ (*true*) means there are expressions with empty domain, and as such, there may be endogenous variables with no parents. Following many others (e.g., Hiddleston 2005, Briggs 2012, Kaufmann 2013) we restrict our attention to **recursive** SEMs, where the parent relation is acyclic.

Intuitively, the expression φ_X specifies the value of X as a function of the variables in $dom(\varphi_X)$ via the *structural equation* $X = \varphi_X$. This intuition is made precise in the following: a **truth assignment for** $\mathscr{M}$ is a function $v\colon \mathsf{Var} \to \{\mathsf{true}, \mathsf{false}\}$ such that for every $X \in \mathsf{End}$, $v(X) = v(\varphi_X)$, where by (a minor) abuse of notation we identify v with its standard recursive extension to all Boolean expressions over Var. Given a Boolean expression φ over Var, we write $\mathscr{M} \Vdash \varphi$ and say that $\mathscr{M}$ **forces** φ just in case for all truth assignments v for $\mathscr{M}$, $v(\varphi) = \mathsf{true}$.

Given $X, Y \in \mathsf{Var}$, we wish to identify those SEMs $\mathscr{M}$ that count as "acceptable explanations" of the counterfactual conditional in (10).

(10) If it had been the case that X, it would have been the case that Y.

As a first pass, let us consider those models $\mathscr{M}$ such that $\mathscr{M} \Vdash X \to Y$. To be sure, the material conditional has long been rejected as a suitable interpretation of natural language conditionals; in this case, however, the forcing relation provides a certain universal character that serves to bring the interpretation more in line with what we might be looking for. More precisely, observe that $\mathscr{M} \Vdash X \to Y$ just in case *every* truth assignment for $\mathscr{M}$ that makes X true also makes Y true; that is, the structural equations in $\mathscr{M}$ guarantee that the truth of X implies the truth of Y. At a high level, this accords with the basic intuition underlying the use of SEMs to interpret counterfactual conditionals: namely, that they encode precisely the kind of antecedent-consequent relationships that licence such utterances.

That being said, there are models that force $X \to Y$ while failing, intuitively, to correspond to explanations for the counterfactual in (10). One problematic case consists in those models $\mathscr{M}$ for which X *cannot* be true under any truth assignment; in this case, $\mathscr{M} \Vdash X \to Y$ holds vacuously. As the impossibility of X seems quite at odds with (10), we reject it as an acceptable explanation, and formally excise it from consideration. Define

$$\mathscr{M} \Vdash X > Y \quad \Leftrightarrow \quad \mathscr{M} \Vdash X \to Y \text{ and } \mathscr{M} \nVdash \neg X.$$

One can readily check that each of the SEMs in Figures 2, 3, 4, 5, and 6 forces $A > \neg B$, which bodes well for this definition. On the other hand, the SEM in Figure 7 also forces $A > \neg B$, despite the fact (discussed in §2.3.5) that this kind of explanation is rejected for standard counterfactual conditionals (i.e., those without additional syntactic licensing, as in (7)).

This leads us to a final refinement. Define

$$\mathscr{M} \Vdash X \rhd Y \quad \Leftrightarrow \quad \mathscr{M} \Vdash X > Y \text{ and } Y \not\prec_{\mathscr{M}} X.$$

We call $\mathscr{M}$ an **acceptable explanation** of (10) provided $\mathscr{M} \Vdash X \rhd Y$. It is easy to see, by this definition, that the SEMs in Figures 2, 3, 4, 5, and 6 are all acceptable explanations of (1), while the SEM in Figure 7 is not.[4]

It is worth noting that the requirement by which we obtained $X \rhd Y$ from $X > Y$—that Y *not* be an ancestor of X—is *not* equivalent to the requirement that X be an ancestor of Y. There are three logical possibilities: either $X \prec Y$, or $Y \prec X$, or neither $X \prec Y$ nor $Y \prec X$. The class of SEMs such that $M \Vdash X \rhd Y$ is thereby naturally partitioned into two components: those where X is an ancestor of Y, and those where neither is an ancestor of the other. A canonical example of an SEM that would lie in the second component is the one pictured in Figure 5, corresponding to the common cause type of explanation. In other words, essentially the same mechanism that excises classical backtrackers from the denotation of a counterfactual also serves to distinguish common cause type explanations from the others.

3.2.1 Extensions

The definition of $X > Y$ is useful as more than just a stepping-stone along the way to the final formulation of acceptable explanations. The SEMs $\mathscr{M}$ that force $X > Y$ but not $X \rhd Y$ have the property that $Y \prec_{\mathscr{M}} X$; they are our classical backtrackers. As we saw in §2.3.5, such explanations become acceptable when the counterfactual is embedded in the right syntactic environment, namely, the double-auxiliary construction. Apparently, in such cases, the requirement that $Y \not\prec X$ is relaxed. We therefore propose to take $X > Y$ as the semantic interpretation of counterfactual conditionals with the double-auxiliary construction.

The definition of $X \rhd Y$ is a strengthening of the definition of $X > Y$; in particular, $X > Y$ does not *exclude* non-backtracking explanations. This turns out to account for two different but related facts about counterfactual conditionals with the double-auxiliary construction: they seem to bias classical backtracking interpretations while still allowing (albeit with some difficulty) non-backtracking interpretations. We can account for the bias by appeal to general Gricean reasoning principles (Grice 1975): if the speaker is going out of her way to use the strictly weaker $X > Y$ rather than $X \rhd Y$, then she must be doing so because the explanations compatible with $X \rhd Y$ are not sufficient. But while this biasing effect makes other interpretations harder to

[4] Technically, $\neg B \not\prec_{\mathscr{M}} A$ is undefined, since $\prec_{\mathscr{M}}$ is a relation on variables, not Boolean expressions. However, to avoid more cumbersome notation, we will suffer this minor abuse, identifying a negated variable with the variable itself for the purposes of assessing ancestorship.

get, they are still available, in accordance with the fact that $X > Y$ is consistent with $X \rhd Y$.

We might also entertain a similar analysis of the *Even...still* construction mentioned in §1. These constructions seem to bias models where the truth of the consequent is fixed. One might argue that models $\mathscr{M}$ satisfying $\mathscr{M} \Vdash Y$ should not count as explanations of (10), and view the *Even...still* construction as relaxing this restriction in just the same way that the double-auxiliary relaxes the $Y \not\prec X$ restriction. In particular, this would imply that the *Even...still* construction biases interpretations where the truth of Y is fixed, which accords well with intuition. However, it is not clear that all the standard interpretations are still available (even with effort) when this construction is used; moreover, some speakers report being able to interpret Y as fixed even without the *Even...still* construction.

3.3 Update

Having formalized the notion of *acceptable explanations* in terms of SEMs, we now describe how to use this framework to capture the effect of asserting counterfactual conditionals vis-à-vis updating the context set. This requires an enriched notion of possible worlds that encode more than just truth values—they must also have something to say about the dependencies between variables. We accomplish this by having worlds encode SEMs.

A **space of structured possible worlds $\mathfrak{M}$ (over** Prop**)** is a nonempty set W together with, for each world $w \in W$, a structural equation model

$$\mathscr{M}_w = (\mathsf{Prop}, \mathsf{End}_w, \{\varphi_{w,X} : X \in \mathsf{End}_w\})$$

and a truth assignment v_w for $\mathscr{M}_w$. For each $X \in \mathsf{Prop}$, we write $(\mathfrak{M}, w) \models X$ just in case $v_w(X) = \mathsf{true}$; this is extended to all Boolean expressions over Prop in the usual way. Thus, each world w encodes the truth or falsity of each Boolean expression φ. But each world also encodes an SEM $\mathscr{M}_w$ specifying dependencies between propositional variables; this can be leveraged to define the truth or falsity of a counterfactual conditional like (10) at a world w as follows:

$$(\mathfrak{M}, w) \models X \rhd Y \quad \Leftrightarrow \quad \mathscr{M}_w \Vdash X \rhd Y.$$

Now if we define the **extension of φ** in the usual way,

$$[\![\varphi]\!] = \{w \in W : (\mathfrak{M}, w) \models \varphi\},$$

then the effect of asserting φ can be captured by updating the context set $C \subseteq W$ with $C \cap [\![\varphi]\!]$, as is standard. In particular, since

$$[\![X \rhd Y]\!] = \{w \in W : \mathscr{M}_w \Vdash X \rhd Y\},$$

we see that an assertion of a counterfactual like (10) eliminates precisely those worlds w where the associated SEM $\mathscr{M}_w$ is not an acceptable explanation of (10).

4 Discussion

4.1 Competing explanations

Recall that the double-auxiliary construction can bias the kind of backtracking explanation shown in Figure 7. What other biases might exist, and what might trigger them? It is entirely possible, for instance, that contextual factors can work to favor some types of explanation over others, or that explanations involving, say, fewer nodes are generally preferred to more complex ones. We leave a systematic study of such biasing factors to future work, but we take a moment here to discuss an issue that lies at the heart of the present enterprise: rejecting explanations.

When might an otherwise acceptable explanation be rejected? For one, we might wish to reject explanations that run counter to well-known regularities in the world. For example, if we know the consequent of a counterfactual to be temporally prior to the antecedent, a dependency of consequent on antecedent is problematic. The counterfactual in (11), for instance, should not be explained by positing a direct dependence of Bob's brunch attendance on Alice's party attendance.

(11) If Alice hadn't come to the party this evening, Bob would have attended the brunch earlier today.

Rather than building this restriction into the notion of an *acceptable explanation*, however, we might treat it as encoded in the common ground, as a background assumption on the part of the conversational participants. Such a move allows us to make a principled distinction between explanations that are ruled out through the update process in virtue of their form (i.e., those that are not *acceptable*, in the technical sense defined in §3.2), and those that merely fail to be considered live possibilities (e.g., no world in the context set encodes an SEM in which $A \prec B$).

Another situation in which we might wish to reject an explanation is when it fails to conform to some *contextually-determined standard of specificity*. This is intended to capture the intuition that we seem to have different expectations for what level of detail is appropriate in different contexts. For example, while taking a Physics test, one might expect that a high level of specificity is required (to demonstrate that one understands the material). By contrast, in a conversation among close friends, shared knowledge might well allow interlocutors to gloss over a great deal of detail; to explicitly spell out more than the minimum might even be insulting. Indeed, we seem to be able to change this *specificity parameter* on the fly, requesting more detail than an interlocutor might have thought was necessary: "Why?" is almost always a felicitous follow-up in the right kind of conversation; one can always raise the bar for the level of detail required.[5]

If a given explanation fails to meet the appropriate standards for specificity in a conversation—either because it is too detailed or not detailed enough—we might wish to reject it. As above, such a rejection seems closely tied to what is represented in the common ground, given that it depends on what is mutually assumed and expected. How best to model this kind of specificity parameter, however—and how to represent conversational moves that change it—we leave for future work.

4.2 Predictions for other languages

Our categorization of the different *types* of explanation was motivated mathematically, not by any particular facts about English counterfactuals, and so it should extend to uses of counterfactuals crosslinguistically. However, there is no *a priori* reason to predict that all languages use a single, underspecified form for all types of acceptable explanations, nor that backtracking explanations should always be differentially marked. A crosslinguistic approach to the modeling of counterfactuals via SEMs promises to be a rich area of investigation.

5 Conclusion

A single counterfactual may be compatible with a multitude of explanations; in this paper we have shown how the SEM framework can be utilized to represent and categorize the event-dependencies relevant to these explanations. This provides a finer-grained analysis of the truth conditions of counterfactual conditions than is present in the extant literature. Moreover, by incorporating SEMs into possible worlds, we are able to bring these insights in line with mainstream compositional semantics: our notion of update is essentially the standard one operating over an enriched space of possible worlds.

The resulting system is relevant not only to the analysis of counterfactuals, but for any linguistic material that is sensitive to the dependencies between events, including non-counterfactual conditionals. Counterfactuals provide an intuitive entry-point into such a system, along with natural illustrations as to the value of building models that encode event dependencies. Applications beyond counterfactuals are left to future work.

[5] Young children often enjoy taking advantage of this conversational move, once they discover it.

References

Asher, Nicholas, Daniel Hardt & Joan Busquets. 2001. Discourse parallelism, ellipsis, and ambiguity. *Journal of Semantics* 18(1). 1–25.

Briggs, Rachael. 2012. Interventionist counterfactuals. *Philosophical Studies* 160(1). 139–166.

Edwards, Brian J & Lance J Rips. 2012. Explanations of counterfactual inferences. In Naomi Miyake, David Peebles & Richard P. Cooper (eds.), *Cognitive Science Society* 34, 318–323.

Grice, H Paul. 1975. Logic and conversation. *Syntax and Semantics* 3. 64–75.

Grice, H Paul. 1978. Further notes on logic and conversation. *Syntax and Semantics* 9. 113–127.

Hiddleston, Eric. 2005. A causal theory of counterfactuals. *Noûs* 39(4). 632–657.

Kaufmann, Stefan. 2013. Causal premise semantics. *Cognitive Science* 37(6). 1136–1170. doi:10.1111/cogs.12063. http://dx.doi.org/10.1111/cogs.12063.

Kratzer, Angelika. 1981. Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic* 10(2). 201–216.

Levinson, Stephen C. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press.

Lewis, David K. 1973. Counterfactuals and comparative possibility. *Journal of Philosophical Logic* 2(4). 418–446.

Lewis, David K. 1979. Counterfactual dependence and time's arrow. *Noûs* 13. 455–476.

Pearl, Judea. 2000. *Causality: Models, Reasoning and Inference*. Cambridge Univ Press.

Rips, Lance J & Brian J Edwards. 2013. Inference and explanation in counterfactual reasoning. *Cognitive Science* 37(6). 1107–1135.

Simons, Mandy. 2012. Conversational implicature. In Claudia Maienborn, Klaus Von Heusinger & Paul Portner (eds.), *Semantics: An International Handbook of Natural Language and Meaning*, vol. 33, 2460–2486. Walter de Gruyter.

Stalnaker, Robert. 1974. Pragmatic presuppositions. In Milton K. Munitz & Peter K. Unger (eds.), *Semantics and Philosophy*, 197–214. New York: New York University Press.

Starr, William B. 2012. Structured possible worlds. Ms. Cornell University.

Wright, Sewall. 1921. Correlation and causation. *Journal of Agricultural Research* 20(7). 557–585.

Adam Bjorndahl
Baker Hall 161
Carnegie Mellon University
Pittsburgh, PA 15213
abjorn@andrew.cmu.edu

Todd Snider
203 Morrill Hall
Cornell University
Ithaca, NY 14850
tns35@cornell.edu