Proceedings of SALT 24: 494–513, 2014

# Filtering semantics for counterfactuals: Bridging causal models and premise semantics *

Paolo Santorio
*University of Leeds*

**Abstract** I argue that classical counterfactual semantics in the style of Stalnaker, Lewis, and Kratzer validates an inference pattern that is disconfirmed in natural language. The solution is to alter the algorithm we use to handle inconsistency in premise sets: rather than checking all maximally consistent fragments of a premise sets, as in Krazter's semantics, we selectively remove some of the premises. The proposed implementation starts from standard premise semantics and involves a new 'filtering' operation that achieves just this removal. The resulting semantics is interestingly related to the semantics for counterfactuals emerging from Judea Pearl's causal models framework in computer science: in particular, filtering is a possible worlds semantics counterpart of Pearl's interventions.

**Keywords:** Counterfactuals, premise semantics, causal models, Kratzer, Pearl

## 1 Introduction

The semantic theories formulated by Stalnaker (1968) and Lewis (1973a, 1973b), together with Kratzer's implementation of these theories (1981a, 1981b, 1986, 1991), are the cornerstone of modern work on counterfactuals, and conditionals in general. While several details have been called into question, the basic structure of these semantic theories—in particular, their reliance on a fixed ordering of worlds that models similarity—has remained untouched. Accordingly, there is wide consensus about what logic is validated by counterfactuals in natural language.

This paper raises a challenge for the classical picture. I show that all existing versions of ordering and premise semantics for counterfactuals validate an inference pattern that is disconfirmed by counterfactuals in natural language. I sketch a new semantics that is able to capture these facts, which I call 'filtering semantics'. Filtering semantics is a close descendant of standard premise semantics for conditionals, but involves an extra 'filtering' operation. Classical premise semantics tests whether

* Thanks to Fabrizio Cariani, Hanti Lin, Alejandro Pérez Carballo, Wolfgang Schwarz, Will Starr, and audiences at the ANU, the University of Sydney, PhLiP 2013, Yale, the University of Leeds, the Philosophy of Language in the UK workshop, the University of East Anglia, and SALT 24.

the antecedent of a conditional, together with the propositions contained in a premise set, entails the consequent. In addition to doing this, filtering semantics selectively removes some information from the premise set. In terms of orderings, the result is equivalent to a semantics that implements a kind of antecedent-driven ordering shift.

Filtering semantics is directly inspired by the causal models framework, a formal framework that has been developed in computer science to give a formal representation of causal relations and causal processes (see, among many, Galles & Pearl 1998 and Pearl 2000). Much recent work has gone into mapping the relationship between causal models and the semantics for counterfactuals; some of this work focuses just on implementing causal models-based ideas in a possible worlds semantics (Schulz 2011, Kaufmann 2013). The theory I defend here is both related and indebted to these accounts, but departs more radically from classical theories. Existing accounts preserve the basic features of classical semantics, including its logic. On the contrary, capturing causal models-type reasoning in possible worlds semantics involves a change of logic. This claims is backed by a recent technical result in Halpern 2013. While I have only learned of Halpern's result after completing the bulk of the present research, this paper can be seen as an exploration of the consequences of his result for the semantics of natural language.

I quickly review standard premise semantics for counterfactuals in section 2 and introduce a puzzle for it in section 3. Then I give a brief and informal introduction to causal models-style reasoning (section 4) and show how this reasoning can be implemented in premise semantics (sections 5 and 6).

## 2 Premise semantics for counterfactuals

### 2.1 Preliminaries: Ordering semantics

Virtually all contemporary accounts of counterfactuals in the possible worlds tradition start from a simple idea, which is pithily put by Stalnaker:

> "Consider a possible world in which $A$ is true, and which otherwise differs minimally from the actual world. "If $A$, then $B$" is true (false) just in case $B$ is true (false) in that possible world." (Stalnaker 1968)

The challenge is explicating rigorously what "differing minimally" amounts to. Accounts in the tradition of Stalnaker and Lewis (1973a, 1973b) do so by appealing to an ordering on worlds. The key formal tool is a relation of comparative closeness, represented as '$\preceq_w$'. $\preceq_w$ compares worlds with respect to their closeness to a benchmark world $w$: $\ulcorner w' \preceq_w w'' \urcorner$ says that $w'$ is closer to $w$ than $w''$ is. The exact way in which $\preceq_w$ figures in the truth conditions for counterfactuals varies across specific versions of the semantics. Here is a version that is often used, and that

strikes a middle ground between Stalnaker and Lewis's own accounts:[1]

> $\ulcorner p \boxdot\!\!\rightarrow q \urcorner$ is true at $w$ just in case all $p$-worlds that are closest according to $\preceq_w$ are $q$-worlds

(Roughly, a world $w$ counts as closest to the actual world just in case there is no world that is closer to @ than $w$ is, according to $\preceq_@$.)

## 2.2 Modal premise semantics

For the purposes of this paper, I take as my benchmark theory not ordering semantics, but rather a premise semantics for counterfactuals derived from the work of Kratzer (1981a, 1981b, 1986, 1991).[2] I have two main reasons. On the one hand, Kratzer's semantics has become something of a standard in the literature on modality. On the other, premise semantics lends itself well to implementing the new account. But let me remind you that, as Lewis (1981) showed, ordering semantics and premise semantics are provably equivalent. Hence nothing substantial hangs on the choice.[3]

For Kratzer, modalized claims in natural language state the existence of a relation between the proposition expressed by the embedded clause (the *prejacent*) and a certain body of information. Consider (1):

(1) David must be the murderer.

On a first pass, (1) states that the proposition that David is the murderer is entailed by a body of information, which Kratzer thinks of as a set of covert premises. All of Kratzer's semantics for modality results from refining this basic idea.

Kratzer postulates the presence of two contextual parameters which jointly determine which propositions are used as premises: the *modal base* and the *ordering source*. Both are functions from worlds to sets of propositions, though for simplicity I will often treat them just as sets of propositions. Modal base and ordering source play distinct theoretical roles. The modal base includes propositions that are, in some

1 These truth conditions incorporate the so-called limit assumption, i.e., the assumption that, for any antecedent, there is a $\preceq_w$-maximal set of antecedent worlds. The limit assumption is controversial, but it makes no difference to my arguments and greatly simplifies my exposition, so I'll make it throughout the paper.

2 While it is standard to use Kratzer's framework nowadays, it should be pointed out that Kratzer's is not the only or even the first premise semantics framework to appear, either in philosophy or formal semantics. For an earlier versions of premise semantics, see Veltman 1976. The basic idea behind premise semantics can be traced back to pre-Lewisian accounts of counterfactuals, like Chisholm 1946 and Goodman 1947.

3 More precisely, *a subtype of ordering semantics* is provably equivalent to premise semantics—namely, the subtype that employs partial orderings. As Lewis (1981) points out, though, there are plausible ways to interpret ordering semantics with total orderings that can be mapped to premise semantics.

relevant sense, settled in the context. The ordering source includes propositions that are used to generate a ranking of worlds along some appropriate dimension. The precise way in which these notions are understood depends on the flavor of the modal. For example, for the case of epistemic modals, the modal base includes propositions that are *known* by some relevant agent, while the propositions in the ordering source involve information about what is *stereotypical* in the context.

While the propositions in the modal base are assumed to be always consistent, this is not so for the propositions in the ordering source. It might be that a number of propositions can be legitimately used to rank worlds along some dimension, but that no single world can satisfy them all. This introduces a problem for the first pass semantics I sketched above. If our premise semantics merely checked whether the premise set entails the prejacent, we would get disastrous results: all necessity claims like (1) would come out trivially false.

Kratzer's fix is quite natural: rather than looking at the logical relations between the prejacent and an inconsistent premise set, we consider all the biggest *consistent fragments* of the premise set. On this new semantics, a necessity claim like (1) states that all the biggest consistent fragments of the premise set entail the prejacent.

This can be formalized as follows. Say that:

> A set of propositions $S$ is a **maximal consistent superset of $S'$ relative to $S''$** iff
>
> (a) $S$ is a superset of $S'$,
>
> (b) $S$ is consistent,
>
> (c) $S$ is formed from $S'$ by adding zero or more propositions from $S''$, and
>
> (d) if any more propositions from $S''$ were added to $S$, $S$ would be inconsistent.[4]

The schematic truth conditions of a modal necessity claim are:[5]

4 Formally:

(a) $S \supseteq S'$;

(b) $\bigcap S \neq \emptyset$;

(c) $(S - S') \subseteq S''$;

(d) $\neg\exists p \in S'' : p \notin S \wedge \bigcap(S \cup \{p\}) \neq \emptyset$.

5 Here, and throughout the rest of the paper, I assume an intentional system, in which interpretation is relativized to a world parameter, as well to a modal base and ordering source parameter. This is just for simplicity; nothing hangs on this assumption.

(2) $[\![\text{must } \phi]\!]^{w,f,g} = 1$ iff, for every maximal consistent superset $S$ of $f(w)$ with respect to $g(w)$, $S \vDash [\![\phi]\!]^{w,f,g}$

Kratzer's technique for handling inconsistent premise sets is just the main feature of Kratzer's apparatus that filtering semantics will call into question.

### 2.3 Counterfactuals in premise semantics

Kratzer treats all conditional statements as modal statements where the *if*-clause works as a restrictor on the domain of quantification. In terms of the apparatus described above, the proposition expressed by the antecedent is added to the modal base. Schematically, these are the resulting truth conditions:

(3) $[\![\text{if } \phi, \text{ would } \psi]\!]^{w,f,g} = 1$ iff, for every maximal consistent superset $S$ of $f(w) \cup \{\phi\}$ with respect to $g(w)$, $S \vDash [\![\psi]\!]^{w,f,g}$

From here, all we need to get an account of counterfactuals is a specification of a modal base and an ordering source that pertain to counterfactual modality. Kratzer's proposal is the following: the modal base starts out empty, while the ordering source maps each world to a set of propositions that are true at that world. It is a difficult and controversial issue *which* true propositions are picked, and arguably one that has never been fully resolved in Kratzer semantics. But my focus in this paper is elsewhere, so I set this issue aside.

## 3 The puzzle: Loop violations

### 3.1 Loop

Consider the following scenario:

> ***Love triangle.*** Andy, Billy, and Charlie are in a love triangle. Billy is pursuing Andy; Charlie is pursuing Billy; and Andy is pursuing Charlie. Each of them is very annoyed by their suitor and wants to avoid them.
>
> There's a party going on and all of them were invited. None of them ended up going, though each of them kept appraised about whether the person they liked was going to be there. An occasion to spend time with the person they liked, without their suitor being there, would have been sufficient for them to go.

In this situation, (4) is judged true, while (5) is judged false, or at least dubious.

(4) If Andy was at the party, Billy would be at the party.

(5) If Billy was at the party, Andy would be at the party.

By symmetry, we get the following set of judgments. These counterfactuals (call them 'forward loop counterfactuals') are judged true:

(4) $A \boxright B$ If Andy was at the party, Billy would be at the party.

(6) $C \boxright A$ If Charlie was at the party, Andy would be at the party.

(7) $B \boxright C$ If Billy was at the party, Charlie would be at the party.

These counterfactuals (call them 'backward loop counterfactuals') are judged oftentimes false, or dubious:

(5) $B \boxright A$ If Billy was at the party, Andy would be at the party.

(8) $A \boxright C$ If Andy was at the party, Charlie would be at the party.

(9) $C \boxright B$ If Charlie was at the party, Billy would be at the party.

The overall situation is summarized in the following table:

| | |
|---|---|
| $A \boxright B$ ✓ | $B \boxright A$ ✗ |
| $C \boxright A$ ✓ | $A \boxright C$ ✗ |
| $B \boxright C$ ✓ | $C \boxright B$ ✗ |

The problem is simple: it is impossible to accommodate this configuration of judgments in existing kinds of ordering or premise semantics. The proof is particularly quick for Stalnaker's ordering semantics, which assumes that the $\preceq_w$ relation is a total order (i.e. all worlds are comparable, and there are no ties: for all $w'$, $w''$, exactly one of $w' \preceq_w w''$ and $w'' \preceq_w w'$ holds). Here it is:

> Since $\preceq_w$ is a total order, there is a unique closest world to $w$ that is an $A$-world, a $B$-world, or a $C$-world. Call this world $w^*$. Without loss of generality, suppose $w^*$ is an $A$-world. Since $A \boxright B$, $w^*$ is also a $B$-world. Since $B \boxright C$, and since $w^*$ is the closest $B$-world, $w^*$ is also a $C$-world. But then, since the (only) closest $A$-world is also a $C$-world, $A \boxright C$ is true. *QED.*

The proofs for other versions of ordering and premise semantics are more involved, but have the same structure. I leave them as an exercise to the reader (who can anyway consult Halpern 2013 for the proof concerning Lewis-style semantics with limit assumption).

Here is a more general way of stating the point. All existing versions of ordering and premise semantics validate the following rule:[6]

$$\textsc{Loop}\quad \frac{\phi \boxrightarrow \psi \quad \psi \boxrightarrow \chi \quad \chi \boxrightarrow \phi}{\phi \boxrightarrow \chi}$$

To my knowledge, LOOP hasn't been discussed in any detail in the literature on condiionals in philosophy or in semantics, though it does appear in the literature on belief revision and nonmonotonic logic: see Kraus, Lehmann & Magidor 1990. My main empirical claim in this paper is that LOOP is in the lot of inference rules that are invalidated by counterfactuals in natural language.[7]

### 3.2 Loop data—brief discussion

One natural reaction to the LOOP data, given its impact on classical counterfactual semantics, is to try to dismiss it. For reasons of space, I must leave a full discussion of this to another occasion (see Santorio 2014a). But here I can rebut two quick attempts at arguing that the challenge can be met easily.

A first line of reply merely questions the facts as I stated them. The claim is that judgments about (4)–(9) are just unclear and that a large number of speakers hedge on them; hence we can't draw any conclusions from them. It is true that some speakers give hedged judgments on at least some of the relevant counterfactuals, and in particular on backward loop counterfactuals (i.e. (5), (8), and (9)). But notice

6 In fact, something stronger is true: LOOP is an instance of a general rule schema, which I call GENERALIZED LOOP (below). All rules that are instances of GENERALIZED LOOP are valid on classical premise semantics.

$$\textsc{Generalized Loop}\quad \frac{\phi_1 \boxrightarrow \phi_2 \quad \phi_2 \boxrightarrow \phi_3 \quad \ldots \quad \phi_{k-1} \boxrightarrow \phi_k \quad \phi_k \boxrightarrow \phi_1}{\phi_1 \boxrightarrow \phi_k}$$

7 Some better known examples are Antecedent Strengthening, Transitivity, and Contraposition; see Lewis 1973a for a discussion, and Burgess 1981 for a classical, in-depth study of a standard counterfactual logic.

that, to get the problem for standard semantics, we don't need the judgments about *any* of the relevant counterfactuals to be clearcut. Judgments about counterfactuals notoriously rely on large amounts of background knowledge. Perhaps the scenario I described is too unspecific to elicit straight true/false judgments. But what matters is that we observe a clear drop in confidence between forward and backward loop counterfactuals. This is enough to challenge standard counterfactual semantics, since the latter predicts that forward loop counterfactuals entail backward loop counterfactuals. This drop in confidence is an extremely robust phenomenon across speakers and very easy to observe.

A second line of reply tries to handle the data via context dependence. Obviously, standard Kratzer semantics can yield the right predictions for all of (4)–(9) by using two different ordering sources, one for forward loop counterfactuals, the other for backward loop counterfactuals. But, equally obviously, the burden is on the proponent of this line to explain why this solution is plausible at all. It seems extraordinary that context should systematically shift just when backward loop counterfactuals are evaluated. In addition, while some context shifts might be justified on the grounds that they help rescue a sentence, in this case the same ordering source that validates forward loop counterfactuals also validates backwards loop counterfactuals. Hence speakers seem to have no reason to pick a different ordering source for backward loop counterfactuals; on the contrary, they would have reason to stick with a unique ordering source throughout.[8]

## 4 Causal models

This section gives a basic overview of the causal models framework. This introduction is very informal and I feel free to pick and choose among pieces of the framework. In particular, I will completely ignore applications of the framework in a probabilistic setting. This leaves out the main use of causal models in the literature, but it allows me to highlight the conceptual core of a causal-models-based treatment

8 Let me also mention a third line of reply. This line tries to explain the data by exploiting an ambiguity in the temporal parameters involved in the antecedents. The suggestion is that, in all of (4)–(9), the time picked out by the antecedent is earlier than the time picked out by the consequent. As a result, the counterfactuals wouldn't really involve the same antecedents, and the apparent counterexample to LOOP would be just the result of an equivocation. The rebuttal here is simply that I have chosen an example where the counterfactuals are not interpreted in this way. For example, (4), repeated below

(4) If Andy was at the party, Billy would be at the party.

does *not* say that, in all situations where Andy is at the party, Billy arrives or is at the party *at some later time*. The order of arrival at the party is left fully open by (4). On its natural interpretation, (4) says that counterfactual situations where Andy is at the party at the time of speech are also situations where Billy is at the party at the time of speech.

of counterfactuals, i.e. the notion of an intervention.

### 4.1 The basic framework

A causal model can be represented as an ordered pair of two elements: $\langle V, E \rangle$. $V$ is a set of *random variables*. A random variable can be thought of as a set of mutually exclusive and jointly exhaustive outcomes for a process: for example, a random variable may represent whether a thermostat is on or off. We can represent this set with familiar tools by using a partition of logical space, i.e. a Groenendijk and Stokhof-style (1984) question denotation. The second element, $E$, is a set of *structural equations*. Structural equations are mathematical equations that state the relations between different values of random variables. For example, a structural equation may state that the answer 'yes' (or, the value '1') to the question whether the thermostat is on correlates with the answer 'yes' (or, the value '1') to the question whether the temperature in a room is above 70 degrees.

It's useful to go through an example in detail. I will use a classical example from Pearl 2000. Readers familiar with it should feel free to skip ahead.

> *The firing squad.* A firing squad is positioned to execute a prisoner. The squad is waiting for a court order. The court issuing the execution order will result in the captain sending a signal to the two members of the squad, X and Y, who will fire and kill the prisoner. The court not issuing the order will result in the captain not sending the signal, the two riflemen not shooting, and the prisoner remaining alive.

Here is a causal model for this scenario:

| **Random variables** | **Structural equations** |
|---|---|
| U: whether the court orders the execution<br>C: whether the captain sends the signal<br>X: whether shooter X shoots<br>Y: whether shooter Y shoots<br>D: whether the prisoner dies | C = U<br>X = C<br>Y = C<br>D = max(X, Y) |

Random variables are traditionally divided into *exogenous* and *endogenous* ones. Exogenous variables are those whose values are determined by factors external to the model. Endogenous variables, conversely, are those whose values are determined by factors within the model. U is the only exogenous variable here.

Strictly speaking, structural equations can be read in either direction, since they are just mathematical equations. But they are conventionally given a 'directional'

reading. The value of the variable on the left-hand side is taken to be determined by the value of the variable on the right-hand side. Hence, for example, 'X = C' is read as indicating that whether rifleman X shoots is determined by whether the captain issues the signal. This directional reading will be crucial in what follows.

In general, in a causal model there is no guarantee that the set of equations will have a unique solution, or any solution at all, for all or even some set of input variables. Oftentimes, though, theorists narrow down consideration to causal models that do have unique solutions in this sense. One important subclass of models that possess this feature is the class of so-called *recursive* models. Recursive models are the ones in which we can define a relation $\prec$ between random variables such that: (a) $X \prec Y$ iff the value of X is *not* dependent on the value of Y; and (b) $\prec$ is a total order. Intuitively, recursive models are the ones where causal dependencies don't go in circles. Our toy model about the prisoner scenario, for example, is a recursive model. Recursive models are not the only models where a unique solution to the equations is available. For example, as I point out below, the Loop-violating scenario I described in section 3 gives rise to a nonrecursive model with a unique solution.

### 4.2 Evaluating counterfactuals

Causal models can be used to provide an evaluation procedure for counterfactuals. The key notion is that of an *intervention*. As a first approximation, an intervention is a manipulation of one of the variables that is made 'from the outside' of a model: i.e., a manipulation that doesn't go through the variables that are causally upstream within the model. Technically, an intervention consists in the replacement of one of the structural equations in the model with a different equation. To evaluate a counterfactual, we proceed in two steps. First, we perform an intervention on the model to make the antecedent true. Then, helping ourselves to the modified set of equations and holding fixed the values of the exogenous variables, we recalculate the values of the endogenous variables. In doing this, we determine whether the consequent holds in the modified model or not. Technically, this means that the evaluation of a counterfactual in a causal model $\langle V, E \rangle$ requires the construction of a derived model $\langle V, E' \rangle$, which involves a modified set of equations.

For illustration, take again the prisoner scenario and suppose that the court didn't issue the execution order. Then all the variables in the model receive value 0 and the prisoner stays alive. Consider the following counterfactual, as uttered against this factual background:

(10) If X had fired, the prisoner would have died.

The first step for evaluating (10) is replacing the old equation with X on the left-hand side with a new equation (in boldface) that specifies the new value of X:

C = U
**X = 1**
Y = C
D = max(X, Y)

At this point, we recalculate from scratch the values of the endogenous variables. From the new equation 'X = 1', together with the equation 'D = max(X, Y)', we get that D = 1, i.e. the prisoner dies. Hence the counterfactual is evaluated as true.

This evaluation procedure is designed to handle a limited range of counterfactuals. Galles & Pearl 1998 and Pearl 2000 restrict themselves to counterfactuals where antecedents are simple sentences—essentially, atomic sentences of the language or conjunctions thereof (though see, among others, Briggs 2012 for an interesting attempt at generalizing the procedure to more complex counterfactuals). Of course, one advantage of implementing this algorithm in a semantics for natural language is that we automatically get a fully general formal system for handling counterfactuals of any complexity.

## 5 Filtering semantics: Basics

### 5.1 Overview

As Kratzer points out, the resolution of inconsistency is one of the central elements in a semantics for counterfactuals:

> "Premise sets can be inconsistent, so the mechanism I was after had to be able to resolve inconsistencies. I believed then, and still believe now, that the semantics of modals and conditionals offers an ideal window into the way the human mind deals with inconsistencies." (2012, p. 1)

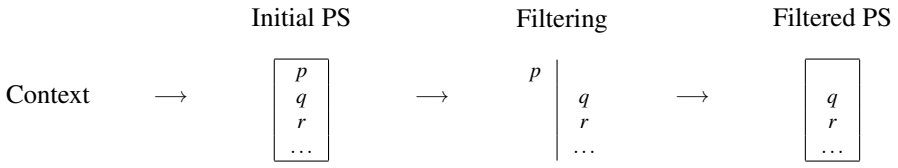
Classical premise semantics handles inconsistent premise sets by considering all maximal consistent subsets of the inconsistent set. Crucially, the causal models-based procedure for evaluating counterfactuals operates in a different way. Together with the inconsistency-generating antecedent, we receive instructions to *remove* some specific piece of information from our previous stock. Hence, together with the *addition* of information to the existing stock, we have a *loss* of previously existing information. This solves immediately the problem of inconsistency; there is no need to consider subsets of the premise set. The rest of this section is devoted to specifying a new premise semantics that implements this conceptual shift.

The main novelty in the semantics is the filtering operation. On classical premise semantics, the antecedent of a counterfactual is simply added to the (otherwise

empty) modal base:

(3) $[\![\text{if } \phi, \text{ would } \psi]\!]^{w,f,g} = 1$ iff, for every maximal consistent superset $S$ of $f(w) \cup \{\phi\}$ with respect to $g(w)$, $S \vDash [\![\psi]\!]^{w,f,g}$

The new semantics adds an extra step: the ordering source is filtered for the antecedent. Hence, while some information is added to the modal base, some other information is removed from the ordering source.[9] In diagram form:




I say that the union of the modal base and the ordering source is *filtered for the antecedent* ($f(w) \cup g(w)$ is filtered for $\phi$). I represent this operation by the vertical bar 'l', using '$X|p$' for '$X$ is filtered for $p$'. On a first pass, the new meaning of counterfactuals can be represented as follows:

(11) $[\![\text{if } \phi, \text{ would } \psi]\!]^{w,f,g} = 1$ iff $[\{\phi\} \cup g(w)]|\phi$ entails $\psi$

Notice one effect of filtering: in general, counterfactuals with different antecedents filter out different information from the ordering source. Hence they are evaluated with respect to different sets of propositions. Premise sets become antecedent-dependent.

## 5.2 Premises and filtering

The implementation of filtering requires modifying the format of the ordering source. Recall from section 4: interventions crucially exploit the directionality of the equations. To implement a similar algorithm in premise semantics, we need to keep track of direction as well—we need to be able to say what determines what. Hence the premises we use need to be more informative than in standard systems.

To this end, I treat the members of the ordering source not as propositions, but as pairs of a Groenendijk and Stokhof question denotation and a proposition. Intuitively, the question specifies which random variable is settled by the proposition. For example, the equation 'X = C' is turned into the pair:

9 Notice: I'm assuming, together with Kratzer herself, that the ordering source of counterfactuals contains consistent propositions, and that the only potential element of inconsistency is generated by the addition of the antecedent to the modal base. This is in line with the common assumption that the ordering source in use for counterfactuals specifies how similar other worlds are to a single world, i.e. the actual world. This ensures that all the propositions that are used to induce the ordering are consistent (since they're all true in the actual world).

> $\langle\{\{w$: $X$ fires in $w\}$, $\{w$: $X$ doesn't fire in $w\}\}$, $\{w$: $X$ fires iff $C$ gives the order in $w\}\rangle$

The question element indicates that the proposition settles whether X fires or not. The proposition element specifies the conditions under which X fires. A *premise* is a pair of a question and a proposition. For simplicity, I take all questions in play to be binary yes-no questions, though this is not required.

The filtering mechanism uses questions to determine which premises should be filtered out by the antecedent. On this basic version of the semantics, a premise is filtered just in case the antecedent settles the answer to its question. The intuition lying behind this is obvious: conditional antecedents are used to settle the answers to questions in the premise set.

Here is a formal statement of the algorithm. Let's say that:

> **A proposition** $p$ **answers a premise** $P$ iff $P = \langle Q, r\rangle$ and $p \in Q$.

With this definition in hand, we can define the filtering of a premise set:

> A **filtering** of a premise set $\Pi$ relative to proposition $p$ is a premise set $\Pi'$ such that, for all premises $P \in \Pi$:
>
> - if $P$ is not answered by $p$, $P \in \Pi'$;
> - if $P$ is answered by $p$, $\langle\{p, \bar{p}\}, p\rangle \in \Pi'$.

In short, we build a filtered premise set $\Pi'$ from an original premise set $\Pi$ by (a) carrying over any premise that is unaffected by the antecedent, and (b) replacing premises whose question is settled by the antecedent with a simple premise where the question consists of the antecedent and its negation, and the proposition is just the antecedent itself.

To state a semantics, we need one further piece of apparatus. Premise sets are now more complex than simple sets of propositions. Hence, as things are, we cannot use the standard notion of a proposition being entailed by a premise set. The fix is simple: we just take the set of all propositions involved in a premise set. I call this the *proposition set* of a premise set $\Pi$, or $\text{Prop}_\Pi$. Formally:

> The **proposition set** of a premise set $\Pi$ is the set $\text{Prop}_\Pi$ such that:
> $\text{Prop}_\Pi = \{p : \exists P \in \Pi : \text{for some } Q, P = \langle Q, p\rangle\}$

Here is a semantics for counterfactuals (minimally different from the preliminary entry in (11)):

(12) $[\![\text{if } \phi, \text{would } \psi]\!]^{w,f,g} = 1$ iff the proposition set of $[\{\phi\} \cup g(w)] | \phi$ entails $\psi$

I discuss an example a few paragraphs below.

### 5.3 Causal ordering sources

The foregoing settles the structural features of the new semantics. But what information is built into the ordering source? This issue is not my main focus in this paper. But it's useful to make some assumptions, so that I can have an account that is able to yield predictions. I assume that ordering sources for counterfactuals (or at least, counterfactuals concerning causal processes) may include the kind of information that is normally included in causal models. This includes: (a) information about causal dependencies and independencies between relevant events and (b) information about some background facts. For illustration, this is how the equations in the execution model get translated into premises:[10]

$$\begin{array}{lcl} \text{C = U} & \Rightarrow & \langle\{c,\bar{c}\},c \leftrightarrow u\rangle \\ \text{X = C} & \Rightarrow & \langle\{x,\bar{x}\},x \leftrightarrow c\rangle \\ \text{Y = C} & \Rightarrow & \langle\{y,\bar{y}\},y \leftrightarrow c\rangle \\ \text{D = max(X, Y)} & \Rightarrow & \langle\{d,\bar{d}\},d \leftrightarrow (x \vee y)\rangle \end{array}$$

Notice that every question in the pairs is related to the random variable appearing on the left-hand side of the equation. Also the background information, i.e. the information about the values of exogenous variables, is encoded in this form. This time the question in play has as its members the proposition itself and its negation. For example, here is how I treat the only exogenous variable in the prisoner scenario (assuming that the court does *not* issue the order):

$$\text{U = 0} \quad \Rightarrow \quad \langle\{u,\bar{u}\},\bar{u}\rangle$$

I should flag that I'm making a number of simplifying assumptions here.[11] In the end, much more work is required to build causal information in the ordering source while tracking Pearl-style dependency relations. (See Kaufmann's framework in his 2013 paper for an attempt at carrying out some of this work.) In any case, this is not

10 For readability, I use some italic letters to stand for the relevant propositions.

11 First, I'm assuming that, for any counterfactual, we can specify an appropriate list of equations and background facts with respect to which the counterfactual is evaluated. Second, I'm assuming that we can appeal to a clearcut distinction between "background" variables, whose causal history we ignore, and "foreground" variables, whose causal history we track via propositions about causal dependencies. This distinction corresponds to the distinction between exogenous and endogenous variables. Third, I'm assuming that, for each context, we can single out a determinate stock of all and only causally relevant variables and dependency relations that we can represent into the ordering source. In short, I'm importing into a Kratzer-style semantics the idealizing assumptions that are required for modeling a situation via a (nonprobabilistic) causal model.

crucial for my purposes in this paper. My main goal here is showing how a causal models-inspired semantics has the logical features we need to capture violations of LOOP.

### 5.4 Accounting for Loop violations

First, let me walk you through a basic example. Consider again (10), repeated below:

(10) If X had fired, the prisoner would have died.

Here is how the semantics handles (10). The initial premise set (on the left) gives rise to the premise set filtered for the antecedent (on the right, changes in boldface):

$$\text{(13)}\qquad \begin{array}{l} \langle \{c, \bar{c}\}, c \leftrightarrow u \rangle \\ \langle \{x, \bar{x}\}, x \leftrightarrow c \rangle \\ \langle \{y, \bar{y}\}, y \leftrightarrow c \rangle \\ \langle \{d, \bar{d}\}, d \leftrightarrow (x \vee y) \rangle \\ \langle \{u, \bar{u}\}, \bar{u} \rangle \end{array} \quad \Longrightarrow \quad \begin{array}{l} \langle \{c, \bar{c}\}, c \leftrightarrow u \rangle \\ \langle \{\boldsymbol{x, \bar{x}}\}, \boldsymbol{x} \rangle \\ \langle \{y, \bar{y}\}, y \leftrightarrow c \rangle \\ \langle \{d, \bar{d}\}, d \leftrightarrow (x \vee y) \rangle \\ \langle \{u, \bar{u}\}, \bar{u} \rangle \end{array}$$

It's easy to check that the propositions in the new premise set entail the consequent, hence the counterfactual is predicted to be true.

Now, let me show how the new semantics accommodates LOOP-violations. Consider again the Love triangle scenario. Here is a simple causal model for it:[12]

| **Random variables** | **Structural equations** |
|---|---|
| A: whether Andy goes to the party | A = (C $\wedge\neg$B) |
| B: whether Billy goes to the party | B = (A $\wedge\neg$C) |
| C: whether Charlie goes to the party | C = (B $\wedge\neg$A) |

Notice that the model is not recursive; causal dependencies *do* run in circles here. At the same time, (a) the model does contain information that is sufficient to determine the values of the relevant variables;[13] and (b) it yields the verdicts that seem intuitive when it is used to evaluate the relevant counterfactuals. To see this, consider how (4) (repeated below) is evaluated:

(4) If Andy was at the party, Billy would be at the party.

12 Of course, this is not the only causal model we might use to represent the scenario. But the important thing is that this is one natural model for the situation, and moreover one that allows us to capture LOOP-violations.

13 In particular, the model has a unique solution in which all the variables have value 0. As Halpern 2013 points out, just nonrecursive models with a unique solution witness the divergence between the logics generated by causal models and those generated by comparative closeness semantics.

By intervening on A, we obtain a derived model with the following equations:

(14)
$$\begin{aligned} &\mathbf{A = 1}\\ &\mathrm{B} = (\mathrm{A} \wedge \neg \mathrm{C})\\ &\mathrm{C} = (\mathrm{B} \wedge \neg \mathrm{A}) \end{aligned}$$

It's easy to check that, in the new model, B must have value 1 and C value 0.

This algorithm for evaluating (4) is reproduced in all relevant ways in filtering semantics. The initial premise set is on the left-hand side; the premise set filtered for the antecedent of (4) on the right:

(15)
$$\begin{array}{lcl} \langle\{a,\bar{a}\}, a \leftrightarrow (c \wedge \neg b)\rangle & & \langle\{\boldsymbol{a},\bar{\boldsymbol{a}}\}, \boldsymbol{a}\rangle \\ \langle\{b,\bar{b}\}, b \leftrightarrow (a \wedge \neg c)\rangle & \Longrightarrow & \langle\{b,\bar{b}\}, b \leftrightarrow (a \wedge \neg c)\rangle \\ \langle\{c,\bar{c}\}, c \leftrightarrow (b \wedge \neg a)\rangle & & \langle\{c,\bar{c}\}, c \leftrightarrow (b \wedge \neg a)\rangle \end{array}$$

Again, it's easy to check that the premises that result from filtering entail that $b$ is true and $c$ is false. Hence (4) is predicted to be true and (8), repeated below, false.

(8) If Andy was at the party, Charlie would be at the party.

Symmetrically, *mutatis mutandis*, for counterfactuals (5) to (9). In summary, filtering semantics predicts that all forward loop counterfactuals are true, and all backward loop counterfactuals are false, validating the pattern of judgments:

$$\begin{array}{ll} A \boxdot\!\!\rightarrow B \checkmark & B \boxdot\!\!\rightarrow A \times \\ C \boxdot\!\!\rightarrow A \checkmark & A \boxdot\!\!\rightarrow C \times \\ B \boxdot\!\!\rightarrow C \checkmark & C \boxdot\!\!\rightarrow B \times \end{array}$$

Notice how this result is achieved. In each case, the contextually provided premise set is the same, i.e. the one specified on the right-hand side in (15). But different antecedents induce different filterings; hence that premise set is manipulated in different ways, depending on the antecedent.

## 6 Filtering semantics: Complications

As I anticipated, this basic account of filtering needs fine-tuning. Here I don't have space for the details, but I can state the problem and the general form of the solution.

The problem is simple: antecedents may be true in multiple ways. I.e., there may be multiple, alternative ways for an antecedent to settle answers to the relevant questions. To see this, consider once more the prisoner scenario and take the counterfactual:

(16) If rifleman X or rifleman Y had shot, the prisoner would have died.

The antecedent of (16) doesn't trigger any filtering. Recall the premise set I've been using:

(17)
$$\begin{array}{l}\langle\{c,\bar{c}\},c\leftrightarrow u\rangle\\ \langle\{x,\bar{x}\},x\leftrightarrow c\rangle\\ \langle\{y,\bar{y}\},y\leftrightarrow c\rangle\\ \langle\{d,\bar{d}\},d\leftrightarrow(x\vee y)\rangle\\ \langle\{u,\bar{u}\},\bar{u}\rangle\end{array}$$

The problem is obvious: there are (at least) two ways to filter the premise set. The antecedent doesn't settle how to do it. Hence the naïve filtering mechanism I considered above would predict that the premise set doesn't change. This it not the result we want.[14]

The fix is pretty straightforward: we consider *all ways to filter the premise set on the basis of the antecedent*. Hence, for the case of (16), we consider the following three filterings, all derived from the premise set in (17):

$$\Longrightarrow\quad\begin{array}{l}\langle\{c,\bar{c}\},c\leftrightarrow u\rangle\\ \langle\{\boldsymbol{x},\bar{\boldsymbol{x}}\},\boldsymbol{x}\rangle\\ \langle\{y,\bar{y}\},y\leftrightarrow c\rangle\\ \langle\{d,\bar{d}\},d\leftrightarrow(x\vee y)\rangle\\ \langle\{u,\bar{u}\},\bar{u}\rangle\end{array}$$

(18)
$$\begin{array}{l}\langle\{c,\bar{c}\},c\leftrightarrow u\rangle\\ \langle\{x,\bar{x}\},x\leftrightarrow c\rangle\\ \langle\{y,\bar{y}\},y\leftrightarrow c\rangle\\ \langle\{d,\bar{d}\},d\leftrightarrow(x\vee y)\rangle\\ \langle\{u,\bar{u}\},\bar{u}\rangle\end{array}\quad\Longrightarrow\quad\begin{array}{l}\langle\{c,\bar{c}\},c\leftrightarrow u\rangle\\ \langle\{x,\bar{x}\},x\leftrightarrow c\rangle\\ \langle\{\boldsymbol{y},\bar{\boldsymbol{y}}\},\boldsymbol{y}\rangle\\ \langle\{d,\bar{d}\},d\leftrightarrow(x\vee y)\rangle\\ \langle\{u,\bar{u}\},\bar{u}\rangle\end{array}$$

$$\Longrightarrow\quad\begin{array}{l}\langle\{c,\bar{c}\},c\leftrightarrow u\rangle\\ \langle\{\boldsymbol{x},\bar{\boldsymbol{x}}\},\boldsymbol{x}\rangle\\ \langle\{\boldsymbol{y},\bar{\boldsymbol{y}}\},\boldsymbol{y}\rangle\\ \langle\{d,\bar{d}\},d\leftrightarrow(x\vee y)\rangle\\ \langle\{u,\bar{u}\},\bar{u}\rangle\end{array}$$

All the premise sets resulting from this procedure are called **permissible filterings** of the original premise sets. Hence, for example, the three premise sets on the left-hand side of (18) are permissible filterings of the original premise set for the

14 With the current setup of the semantics, we would get back an inconsistent premise set, which would make all counterfactuals with the same antecedent as (16) trivially true (or, if we tried to enforce a kind of nonvacuousness presupposition, defective).

antecedent $\ulcorner x \vee y \urcorner$.

The procedure for determining permissible filterings is something like the converse of the procedure we used to determine what was filtered out in the basic semantics. Then we checked whether the antecedent of a conditional settled the answer to any relevant questions in the premise set. Now we check which answers or combinations of answers in the premise set entail the antecedent of a conditional. The reader may consult Santorio 2014b for a formal statement of this procedure.

The semantics for counterfactuals is modified to quantify over permissible filterings. Here is the new schematic entry:

(19) $[\![\text{if } \phi, \text{ would } \psi]\!]^{w,f,g} = 1$ iff for all $\Pi$ s.t. $\Pi$ is a permissible filtering of $[\{\phi\} \cup g(w)]$ for $\phi$, the proposition set of $\Pi$ entails $\psi$

The new semantics quantifies again over multiple premise sets, as Kratzer's original semantics. But the basic contrast between the two accounts still stands. Kratzer-style semantics considers all consistent fragments of the initial premise sets. Filtering semantics only considers premise sets that we obtain by making changes in the causal path that is upstream with respect to the antecedent.

## 7 Conclusion

My starting point has been a challenge for standard semantics for counterfactuals. All versions of ordering and premise semantics validate an inference pattern, LOOP, that seems disconfirmed in natural language. LOOP-invalidating data deserves to be taken seriously, hence it's worth investigating variants of classical semantics that are able to predict it. Interestingly, one semantics that yields the right predictions in LOOP cases is based on the causal models framework. The key change is that the new semantics, in addition to adding the antecedent to the premise set, also removes some other information from it. This removal process, which I've called 'filtering', results in a different strategy for resolving inconsistency from the one currently used by standard modal and counterfactual semantics.

While the LOOP-invalidating data might not be sufficient to justify a paradigm shift, let me close by flagging some areas of research which may lend further support to filtering semantics: counterfactuals that track causal rather than temporal dependencies (see, among many, Slote 1978, Hiddleston 2005, Ippolito 2013); the treatment of disjunctive antecedents (Fine 1975, Nute 1975); a puzzle for Lewis-style counterfactual logic recently discovered by Kit Fine (2012a, 2012b).

## References


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

Burgess, John P. 1981. Quick completeness proofs for some logics of conditionals. *Notre Dame Journal of Formal Logic* 22(1). 76–84.

Chisholm, Roderick M. 1946. The contrary-to-fact conditional. *Mind* 55(219). 289–307.

Fine, Kit. 1975. Review of Lewis' counterfactuals. *Mind* 84. 451–458.

Fine, Kit. 2012a. Counterfactuals without possible worlds. *Journal of Philosophy* 109(3). 221–246.

Fine, Kit. 2012b. A difficulty for the possible worlds analysis of counterfactuals. *Synthese* 189(1). 29–57.

Galles, David & Judea Pearl. 1998. An axiomatic characterization of causal counterfactuals. *Foundations of Science* 3(1). 151–182.

Goodman, Nelson. 1947. The problem of counterfactual conditionals. *The Journal of Philosophy* 44(5). 113–128.

Groenendijk, Jerome & Martin Stokhof. 1984. *Studies in the semantics of questions and the pragmatics of answers*: University of Amsterdam PhD dissertation.

Halpern, Joseph Y. 2013. From causal models to counterfactual structures. *Review of Symbolic Logic* 6(2). 305–322.

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

Ippolito, Michela. 2013. Counterfactuals and conditional questions under discussion. In Todd Snider (ed.), *Proceedings of Semantics and Linguistic Theory (SALT) 23*, 194–211. http://elanguage.net/journals/salt/index: CLC Publications.

Kaufmann, Stefan. 2013. Causal premise semantics. *Cognitive science* 37(6). 1136–1170.

Kratzer, Angelika. 1981a. The notional category of modality. In Hans-Jürgen Eikmeyer & Hannes Rieser (eds.), *Words, Worlds, and Contexts: New Approaches to Word Semantics*, Berlin: de Gruyter.

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

Kratzer, Angelika. 1986. Conditionals. In Farley Peter T. Farley, Anne M. & Karl-Eric McCullough (eds.), *Chicago Linguistics Society: Papers from the Parasession on Pragmatics and Grammatical Theory*, vol. 22 2, 1–15. University of Chicago, Chicago IL: Chicago Linguistic Society.

Kratzer, Angelika. 1991. Modality. In *Semantics: An International Handbook of Contemporary Research*, 639–650.

Kratzer, Angelika. 2012. *Modals and Conditionals: New and Revised Perspectives*, vol. 36. Oxford University Press.

Kraus, Sarit, Daniel Lehmann & Menachem Magidor. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence* 44(1). 167–207.

Lewis, David K. 1973a. *Counterfactuals*. Cambridge, MA: Harvard University Press.

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

Lewis, David K. 1981. Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic* 10(2). 217–234.

Nute, Donald. 1975. Counterfactuals and the similarity of words. *The Journal of Philosophy* 72(21). 773–778.

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

Santorio, Paolo. 2014a. Causal models and counterfactual logic: Experimental data. In preparation.

Santorio, Paolo. 2014b. Interventions in premise semantics. Manuscript, University of Leeds.

Schulz, Katrin. 2011. If you'd wiggled A, then B would've changed. *Synthese* 179(2). 239–251.

Slote, Michael A. 1978. Time in counterfactuals. *The Philosophical Review* 87(1). 3–27.

Stalnaker, Robert. 1968. A theory of conditionals. In Nicholas Rescher (ed.), *Studies in Logical Theory*, Oxford.

Veltman, Frank. 1976. Prejudices, presuppositions, and the theory of counterfactuals. In *Amsterdam Papers in Formal Grammar. 1st Amsterdam Colloquium*, 248–281. University of Amsterdam.

Paolo Santorio
School of Philosophy, Religion, and History of Science
University of Leeds
Woodhouse Lane
Leeds LS2 9JT
United Kingdom
paolosantorio@gmail.com