
White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/101392/

Version: Accepted Version

Article:
Santorio, P (2019) Interventions in Premise Semantics. Philosophers' Imprint, 19 (1). pp. 1-27. ISSN: 1533-628X

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



Interventions in Premise Semantics

Paolo Santorio
UNIVERSITY OF LEEDS

Final version. Forthcoming in *Philosophers' Imprint*.

1 Introduction

Counterfactual thought and talk play a central role throughout science and in many areas of philosophy. Hence it's not surprising that they have been extensively studied in a number fields: besides philosophy itself, linguistics, psychology, and computer science. In particular, much work has gone into developing a logic and a semantics for counterfactual conditionals, i.e. conditionals of the form:

(1) If Mary pulled the trigger, her gun would fire.

This paper investigates what happens when we merge two different lines of theorizing about counterfactuals, with particular attention to the goal of giving a compositional semantics. One of these views is the comparative closeness view, which was initially developed by Stalnaker (1968) and Lewis (1973a, 1973b) in the framework of possible worlds semantics. The second is the interventionist view, which is part of the causal models framework developed in statistics and computer science (see e.g. Pearl 2000). Common lore (and existing literature) have it that the two views can be easily fit together, aside perhaps from a few details. I argue that, on the contrary, transplanting causal-models-inspired ideas in a possible worlds framework yields a substantially new semantics, which makes systematically different predictions and generates a new logic. The difference is ultimately grounded in different algorithms for handling inconsistent information. Hence it touches on issues that are at the very heart of a semantics for conditionals that involve contrary-to-fact suppositions. The upshot is that we have a new semantics to study, and a substantial choice to make.

The bulk of this paper is devoted to explaining in detail the new view, but it's helpful to give a rough sketch here. Start from classical possible worlds semantics for counterfactuals. Under one popular implementation (so-called premise semantics), here is how the evaluation of a counterfactual works. We hold fixed a set S of true propositions, which work as covert premises, and we check whether those propositions, together with the antecedent, entail the consequent. Schematically:

Thanks to Fabrizio Cariani, Jennifer Carr, Alejandro Pérez Carballo, Wolfgang Schwarz, Will Starr, two anonymous referees at Philosophers' Imprint, and audiences at the ANU, the University of Sydney, PhLiP 2013, Yale University, the University of Leeds, the Philosophy of Language in the UK workshop, the University of East Anglia, SALT 24, and Carnegie Mellon University. Special thanks to Hanti Lin, with whom I have had very fruitful exchanges, in person and via email, about causal models and counterfactuals.

$\ulcorner p \mathrel{\Box\!\!\rightarrow} q \urcorner$ is true iff p, together with propositions in set S, entail q

The new semantics adds an extra step. Rather than using a fixed stock of propositions, we use the antecedent to selectively eliminate some of those propositions from the set. I say that, when this happens, the set S is *filtered for* the antecedent. Accordingly, I call the new semantics *filtering semantics*. Here are the new schematic truth conditions:

$\ulcorner p \mathrel{\Box\!\!\rightarrow} q \urcorner$ is true iff p, together with propositions in set S *filtered for* p, entail q

One alternative version of possible worlds semantics exploits, rather than covert premises, a relation of comparative closeness between worlds. Within this framework, filtering amounts to an antecedent-driven shift in what worlds count as closer by or further away—something that is not contemplated by any standard counterfactual semantics.

Recently, some interesting work has gone into understanding the relationship between the causal models framework and comparative closeness semantics: for example, Schulz 2011 and Kaufmann 2013. The theory I present 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. I argue that, on the contrary, the causal models framework involves a different conception of counterfactual supposition; adopting this conception requires substantial changes to the semantics, as well as a change of logic. This claim 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.[1]

This paper focuses on the positive task of constructing a causal models-based semantics and explaining how it differs from classical counterfactual semantics. Settling which of the two theories is empirically correct goes beyond my purposes—indeed, this doesn't seem the kind of question that can be settled in the space of a single paper. But I can start bringing up some evidence that can decide between the two views. On a preliminary survey, this evidence seems to support filtering semantics against the classical account. This is obviously not enough to justify a paradigm shift, but it shows that filtering semantics deserves more in-depth investigation, and that causal-models-style reasoning should be taken seriously not only by philosophers of science, but also by philosophers of language, semanticists, and logicians.

I quickly review standard semantics for counterfactuals in §2 and introduce the causal models framework in §3. I give a premise semantics that implements causal-models-style reasoning in §4 and §5, and I show how the new semantics diverges in predictions from classical premise semantics in §6.

[1] I should note that at least another philosopher has claimed that, when properly developed, the Pearl framework forces us to depart from standard counterfactual logics: see Briggs 2012. (Briggs, though, seems to suggest that a semantics that yields the new logic should depart entirely from semantics in the possible worlds tradition.) Unfortunately, for reasons of space I can't discuss her specific claims in this paper.

2 Premise semantics for counterfactuals

2.1 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:

> $\ulcorner$If ϕ, would $\psi\urcorner$ is true at w just in case all ϕ-worlds that are closest according to $\preceq_w$ are ψ-worlds

(Roughly, a world w' counts as a closest world to w just in case there is no world that is closer to w than w' is, according to $\preceq_w$.) These truth conditions rely on 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 will adopt it throughout the paper.

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. Nothing substantial hangs on this choice: as Lewis 1981 points out, it is possible to derive orderings from Kratzerian premise sets. Hence the semantics I give could be restated in terms of orderings.[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 (2):

[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] Lewis's paper is sometimes taken to show that premise semantics in the style of Kratzer is equivalent to a particular version of ordering semantics, i.e. the version that uses partial orderings (i.e., roughly, the version of ordering semantics developed and defended by Pollock 1976). But, as I point out in Santorio 2016, this is an overstatement—the equivalence holds only for unembedded counterfactuals and fails for nested counterfactuals. The failure of the general claim concerns differences in the compositional handling of domain restriction and is orthogonal to the main issues I deal with here.

(2) David must be the murderer.

On a first pass, (2) 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 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 (2) 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 (2) states that all the biggest consistent fragments of the premise set entail the prejacent.

Let me start introducing some formalism. Following standard semantic theories, I use an interpretation function (normally represented via the double square brackets '$[\![\cdot]\!]$') to specify a mapping of expressions to compositional semantic values. This mapping is normally relativized to a number of parameters. For current purposes, I take these parameters to include a possible world (represented as 'w'), a modal base (represented as 'f'), and an ordering source (represented as 'g'). So the general form of a semantic clause will be:

$[\![\phi]\!]^{w,f,g}$ = semantic value of ϕ relative to parameters w, f, g

It is customary to relativize interpretation to a context as well, but I will skip this to avoid clutter.

In addition, it's useful to have some quick notation to denote the possible worlds proposition expressed by a sentence (holding fixed a choice of a modal base and an ordering source). I will use the double straight brackets '$\|\cdot\|$' for this purpose. Assuming that a possible worlds proposition is just with a set of worlds, I write:

$$\|\phi\|_{f,g} = \{w : [\![\phi]\!]^{w,f,g} = 1\}$$

Now I can state a basic version of Kratzer semantics for modals in a formal fashion. 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]

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

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 Modal premise semantics: counterfactuals

For Kratzer, all conditional statements are modal statements of sort. The *if*-clause is used, in addition to the modal base, to restrict the domain of quantification of the relevant modal. This is implemented simply by adding the proposition expressed by the antecedent to the modal base. Schematically, these are the resulting truth conditions:[6]

(4) $[\![\text{If } \phi, \text{ would } \psi]\!]^{w,f,g} = 1$ iff, for all maximal consistent supersets S of $f(w) \cup \{\|\phi\|_{f,g}\}$ with respect to $g(w)$, $S \vDash \|\psi\|_{f,g}$

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 this: 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, but set that aside for a moment.) Hence, in Kratzer's apparatus, the ordering source plays the role of orderings in ordering semantics.

Notice that this semantics, as I have stated it, incorporates a version of the limit assumption. In present terms, the assumption is that, no matter how we extend the modal base by adding propositions from the ordering source, we always hit on a maximal consistent superset, i.e. one that cannot be further extended without falling into inconsistency.

---

[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] The notion of entailment between a set of possible worlds propositions S and a proposition p is defined in the obvious way: S entails p iff the intersection of the propositions in S is a subset of p.

$S \vDash p$ iff $\bigcap S \subseteq p$

[6] The semantics in (4) is not designed to handle nested conditionals. To handle embeddings of this sort, we should complicate the semantics by letting the modal base with respect to which the consequent is evaluated be updated by the antecedent. Schematically:

(i) $[\![\text{If } \phi, \text{ would } \psi]\!]^{w,f,g} = 1$ iff, for all maximal consistent supersets S of $f(w) \cup \{\|\phi\|_{f,g}\}$ with respect to $g(w)$, $S \vDash \|\psi\|_{f+\phi,g}$
(where $f + \phi = \lambda w.\, f(w) \cup \|\phi\|_{f,g}$)

Since my argument doesn't touch on nested conditionals, I ignore this complication throughout the paper.

2.4 Causal dependencies and premise semantics

I close my overview by rehearsing a well-known line of argument to the effect that a notion of dependence, and in particular causal dependence, plays an important role in the semantics of counterfactuals.

Let me emphasize that everything I say in this section is compatible with standard premise semantics. The upshot of the argument is that the premise sets we use to evaluate counterfactuals must encode information about causal dependence. This is fully compatible with the mechanics that I described in §2.1–2.3 staying untouched. In premise semantics, questions about the structural and logical properties of modals are orthogonal to questions pertaining to the choice of premise sets. Nevertheless, reviewing the argument helps motivate the shift of attention to notions of dependence, and causal dependence in particular.

To start, consider the following scenario:

> *Coin toss.* Alice is about to toss a coin and offers Bob a bet on heads; Bob declines. Alice tosses the coin, which does indeed land heads.

And now consider the following counterfactual, as evaluated in the above scenario:

(5) If Bob had taken the bet, he would have won.

(5) is judged true. Now, notice what kind of information we need to hold fixed to vindicate this judgment. The moment at which Bob takes or rejects the bet precedes the moment of the coin toss. Moreover, coin tosses (suppose) are indeterministic events. Hence, when Bob takes or rejects the bet, it is indeterminate whether he will win or lose. (If you're doubtful that coin tosses are genuinely indeterministic, just tweak the example.) Hence, if we held fixed only the events in history that preceded Bob's decision, and the proposition that he accepted the bet, it wouldn't be settled that he would win. It might still be that the coin lands tails and he loses. By contrast, we do get the right prediction if we decide to hold fixed all facts that are causally independent of Bob's taking the bet. The actual outcome of the coin toss is independent of Bob's taking the bet; hence we can hold that outcome fixed and use it to generate the conclusion that Bob would have won.

Cases like (5) have been in the literature since Slote 1978 (who credits Sydney Morgenbesser for first introducing them). They point to the idea that natural language counterfactuals track relationships of causal dependence and independence, suggesting that this information should

be incorporated into premise sets and orderings.[7,8] Here I take on board this idea, though I develop it in ways that go beyond the suggestion that orderings and premise sets should incorporate causal information.

Throughout the paper, I will focus on dependencies of a causal kind. But let me flag that, in principle, the notions of dependence in play may be understood in a broader way. Many run-of-the-mill counterfactuals track dependencies of a noncausal nature. For example:[9]

(6) If I had arrived at 2:05, I would have been five minutes late.

My arriving at 2:05 when I have a 2 pm appointment is not cause of my being late, but it somehow determines my being late. Conversely, my being five minutes late depends (in part) on my arriving at 2:05, though it is not caused by it. If we decide to track causal dependencies in the semantics, it seems plausible that we should also track noncausal dependencies of this sort. I will not pursue this extension of the project here. But the formalism that I develop in the next sections is flexible, and in principle there is no obstacle to extending it for modeling all sorts of dependence.[10]

[7] In fact, just examples of this kind have been used to construct a battery of counterexamples to Lewis's stated criteria for determining similarity. Lewis's (1979) criteria are, roughly: (1) avoidance of major violations of actual laws; (2) maximization of spatio-temporal regions in which there is perfect overlap of particular facts with the actual world; (3) avoidance of minor violations of actual laws; (4) in *some* cases (more on this in a minute), vindication of approximate similarity of particular facts with the actual world. Lewis's account is equipped to deal with the case I present in the main text (via clause (4) of the account), but there are very similar cases that are problematic. Here is one that mixes the coin toss with Lewis's own famous Nixon example in (1973a), due to Hiddleston 2005:

> *Chancy nuclear war.* Alice is about to toss a coin and offers to Bob to bet. Unbeknownst to them, Nixon is watching them play. He has decided that he'll push the button to launch a nuclear attack just in case Bob wins the bet. Bob bets on tails. Alice tosses the coin, which lands on heads. Nixon puts away the button.

And now, consider the following counterfactual:

(i) If Bob had bet on heads, he would have won and Nixon would have launched a nuclear attack.

(i) seems true, but this is not predicted by Lewis's metric. Consider worlds where Bob bets on heads and loses because of the coin landing differently, and compare them to the worlds where Bob bets and wins, and there is indeed a nuclear holocaust. The two kinds of worlds are tied with regard to criteria (1)–(3), and worlds of the former kind come out ahead on criterion (4)—assuming that (i) is one of the cases where approximate similarity of particular fact matters. Of course, it is open to Lewis to claim that approximate similarity matters for the case I present in the main text, but not for (i). But this disanalogy does demand for an explanation as Lewis himself acknowledges in "Counterfactual Dependence and Time's Arrow":

> It is a good question whether approximate similarities of particular fact should have little weight or none. Different cases come out differently, and I would like to know why. (Lewis 1979, p. 472)

[8] What about Kratzer? Her most recent account of the information contained in premise sets (which dates back to her 1989) exploits a relationship of 'lumping', i.e. a kind of mereological relationship between situations. It is just unclear to me to what extent this account would manage to incorporate facts about causal dependence and independence.

[9] Thanks to Wolfgang Schwarz for the example.

[10] One may worry here that widening the scope of the semantics in this way makes it unexplanatory. We have an intuitive grip on causal structures that is independent of our judgments about counterfactuals. But we don't have an equally intuitive grip on noncausal dependencies. Hence, in order to build noncausal dependence models, we have to rely on our judgments about counterfactuals, and this will make the semantics uninformative, or unexplanatory, or both. I agree that decisions about dependence modeling may be driven just by our judgments about counterfactuals. But I think that the resulting semantics would still be informative, and even predictive, in key respects. In particular, one central task of a semantics for conditionals is characterizing their logic. This allows us to predict the linguistic data that is the basis of our theorizing about conditionals, i.e. what patterns of conditionals speakers find

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

3.1 The basic framework

The main ambition of the causal models framework is modeling how events in a causal network are dependent or independent of one another, and how a change in the outcome of one event affects the others. While there are different ways of setting up causal models formally, they all rely on the the core ideas I introduce here.

A causal model consists in 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 the state of a thermostat; the thermostat being on and the thermostat being off are the two values of the variable. In philosophy and semantics, this structure is familiar from partitions of logical space, and is often used to capture the denotation of an interrogative clause (see, among many, Lewis 1982, and 1988, and Groenendijk & Stokhof 1984). Hence one intuitive way to think of a random variable is to identify it with the content of a question.

The second element of a causal model, 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:

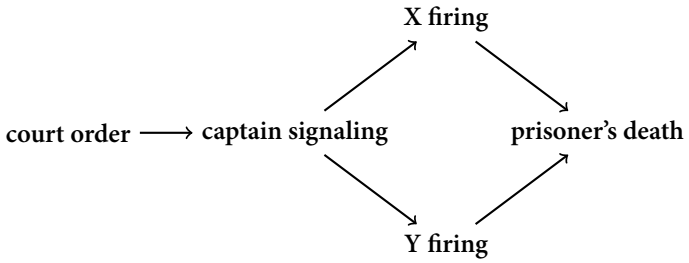
consistent or inconsistent. But what logic we get is independent of choices of background information. Hence it may be that a semantics based on generalized dependence models will not be explanatory at the level of characterizing truth conditions of individual counterfactuals. But it will still provide interesting explanations of the consistency or inconsistency of patterns of counterfactuals, and will be part of an explanatory theory of counterfactual reasoning.

Random variables	Structural equations
U: whether the court orders the execution C: whether the captain sends the signal X: whether shooter X shoots Y: whether shooter Y shoots D: whether the prisoner dies	C = U X = C Y = C 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. In the toy model provided, U is the only exogenous variable. This can be seen from the fact that there is no equation that has U on the left-hand side.

Strictly speaking, structural equations are just mathematical equations and hence can be read in either direction. But, by convention, they are read directionally. 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. Of course, this is more informative than the material equivalence 'X shoots just in case the captain issues the order'. This feature of structural equations is crucial both for capturing counterfactual reasoning within the causal models framework and for the innovations introduced by filtering semantics.

Causal models are usually represented visually by means of directed graphs, i.e. diagrams in which nodes represent random variables and arrows represent relationships of causal dependence. This is the graph corresponding to our toy model:[11]



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. But we can narrow down consideration to classes of 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

[11]Notice that the visual representation generally produces a loss of information. The arrows represent causal dependence, but they are silent about exactly what that dependence involves. For example, the graph above doesn't specify whether the dependence between D, on the one hand, and X and Y, on the other, is conjunctive or disjunctive. it is compatible with two distinct equations having D on the left-hand side:

D = min(X, Y)
D = max(X, Y)

Hence the reader should take graphs just as convenient props. The full specification of a causal model is given by the set of random variables and the set of equations.

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. Graphically, recursive models can be represented via *acyclic* graphs—graphs where one cannot start from and come back to the same point by following the arrows. It should be easy to check that our model about the prisoner scenario is recursive.

Interestingly, recursive models are not the only models where a unique solution to the equation is available. An example involving a nonrecursive model with a unique solution will be at the center of my discussion in §6.

3.2 Evaluating counterfactuals

Causal models can be used to provide an evaluation procedure for counterfactuals. For a number of reasons, this evaluation procedure cannot be seen as a real semantics for counterfactuals in natural language.[12] Nevertheless, some conceptual tools involved in this procedure may be put to use in a compositional semantics.

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 and check whether the consequent holds. Technically, this means that the evaluation of a counterfactual in a causal model $\langle V, E\rangle$ requires building a *derived model* $\langle V, E'\rangle$, which involves a modified set of equations. The derived model is used to assess the consequent for truth and falsity.

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:

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

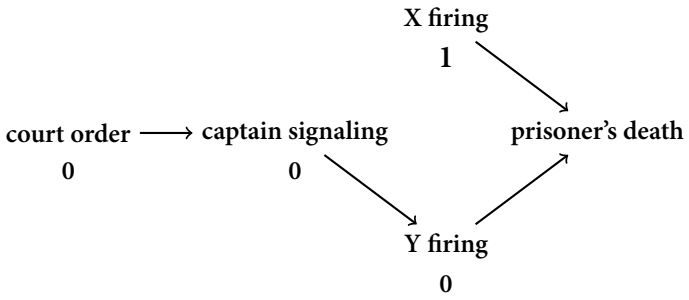
The first step for the evaluation of (7) is the replacement of the old equation with X on the left-hand side with a new equation that specifies the new value for X. Here is the new model:

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

[12] Let me mention here three reasons. First, the procedure provides a way of assigning a truth value to a counterfactual *relative to a causal model*; but, to get a full semantics for counterfactuals out of this, we should be able to say what model is relevant for the evaluation of a counterfactual. For a discussion of this point, see Hiddleston 2005. Second, in discussions of causal models in computer science and philosophy of science there is just no attempt at incorporating the evaluation procedure into a full-blown compositional semantics for natural language. Third, the original procedure formulated by Pearl only covered a subclass of counterfactuals with certain syntactic properties (essentially, counterfactuals whose antecedents involved atomic sentences or conjunctions thereof). While substantial work has been done to broaden this coverage (see for example Halpern 2000 and Briggs 2012), to date there is no comprehensive treatment of counterfactuals of arbitrary syntactic complexity in Pearl's framework.

At this point, holding fixed the values for the exogenous variables, 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. Notice that, given the way that the procedure is set up, all the values of variables that are upstream with respect to the intervention are guaranteed to remain the same; values of other variables may change.

The modified model can be captured by a new graph. The fact that the causal dependence of X on C is now ignored is represented by the fact that the arrow going from the latter node to the former is removed:



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). One advantage of implementing this algorithm in filtering semantics is that we automatically get a general formal system for handling counterfactuals of any complexity.

4 Filtering semantics for counterfactuals: basics

4.1 The goal

Much work has gone into comparing closeness semantics with the interventionist account. Initially, this work has focused on the logic. Galles & Pearl 1998 argued that, if we restrict consideration to recursive models and to a simple language involving exclusively atomic sentences and conjunction, the causal models framework validates the logic generated by Lewis-style possible worlds models (in addition to enforcing some further conditions).[13] Some theorists have also investigated analogous claims for larger categories of models and more expressive languages. The general conclusion (see, for example, Briggs 2012) seems to be that, as soon as we relax the constraints assumed by Galles and Pearl, the logics generated by the causal models framework starts diverging substantially from counterfactual logics in the possible worlds tradition.

[13] I refrain from a more precise statement of this claim, since it is the object of contention: see Halpern 2013: pages 305-307 for discussion. I say 'argued' rather than 'showed' because Galles and Pearl's result, though correct, was obtained via a not entirely correct proof, as Halpern shows.

Attempts at implementing causal-models-type reasoning in a compositional semantics for modality are more recent. The most detailed attempt in this direction is Stefan Kaufmann's (2013). Kaufmann also starts from Kratzer semantics, which he modifies by imposing a lexicographic ordering on the propositions in the premise set. Despite the changes, his conclusion is that implementing the conceptual tools of causal models into premise semantics can be done without major changes:

> [A] premise semantic account of the causal inferences that tend to enter the interpretation of counterfactuals is not only possible, but in fact fairly straightforward. (2013, p. 1163)

It's true that, in the kind of causal scenarios that Kaufmann considers, a modified version of Kratzer's premise semantics will get results and predictions that parallel those of Pearl's framework. But this similarity hides a substantial divergence between the two frameworks, which (as it happens for the logic) is revealed when we start looking at different kinds of cases. This divergences are not a mere accident, but they stem from a conceptual difference between the causal models framework and classical premise semantics: the two frameworks rely on different algorithms for resolving inconsistency. Hence their divergence has to do with the very core of a semantics for counterfactuals. Exposing this difference, and building a causal-models-based semantics that captures it, is my main goal in this paper.

Here is the plan from now on. In this section, I give a basic version of the semantics, highlighting from the start the main element of divergence, and how this element leads to a new semantics. In §5, I introduce some needed refinements, and in §6 I show how the old and the new semantics differ in predictions.

4.2 Overview of the semantics

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

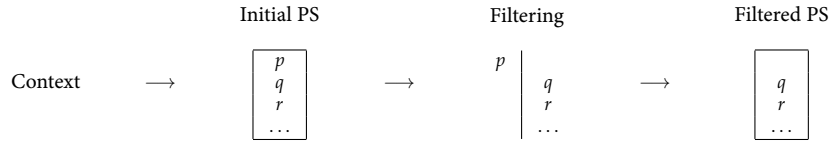
> Premise sets can be inconsistent, so the mechanism I was after had to be able to resolve inconsistencies. I believed then, and still believe now, that the semantics of modals and conditionals offers an ideal window into the way the human mind deals with inconsistencies. (2012, p. 1)

Classical premise semantics handles inconsistent premise sets by considering all maximal consistent subsets of the inconsistent set. Crucially, the causal-models-based evaluation of 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. In this section, I set up a basic version of a new premise semantics that implements this conceptual shift. The next section is devoted to developing a new, more general version of the semantics.

The main innovation is the filtering operation. On classical premise semantics, recall, the antecedent of a counterfactual is simply added to the (otherwise empty) modal base:

(4) $[\![\text{If } \phi, \text{ would } \psi]\!]^{w,f,g} = 1$ iff, for all maximal consistent supersets S of $f(w) \cup \{\|\phi\|_{f,g}\}$ with respect to $g(w)$, $S \vDash \|\psi\|_{f,g}$

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.[14] In diagram form:



I say that the union of the ordering source is *filtered for the antecedent* ($g(w)$ is filtered for ϕ)[15]. I represent this operation by the vertical bar '|', using '$X|p$' for 'X is filtered for p'.

Below is a first-pass new meaning for counterfactuals. As is evident from the entry, the new meaning doesn't require appeal to a modal base; we can do everything with the ordering source (more about this in §5.3).

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

More informally:

> ⌜if ϕ, would ψ⌝, evaluated relative to ordering source g and world w, is true iff the premise set $g(w)$, filtered for ϕ, entails ψ.

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. In other words, the premise sets we use to evaluate consequents become antecedent-dependent.

4.3 Directional premises

The implementation of filtering requires modifying the format of the ordering source. Recall from §3: 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 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:

[14]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).

[15]I will be sloppy with notation, and I will treat filtering as performed equivalently using a sentence, or by a proposition. The official version of the theory employs propositions. The term 'filtering' (as well as the '|' notation) already appears in Cariani et al. 2013. Cariani, Kaufmann, and Kaufmann's filtering is different from mine, though there are interesting formal analogies between their account of deontic modals and my account of counterfactuals. Unfortunately, I cannot explore these analogies here.

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

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 foregoing settles the structural features connected to the ordering source in the new semantics. But what information is built into the ordering source? For present purposes, I am just going to show how to import information from a causal model into an ordering source in premise semantics. While a plausible semantics for counterfactuals might demand more sophistication, I only aim at setting up a basic apparatus that implements causal-models-style reasoning in premise semantics.

The new ordering source, mirroring causal models, will incorporate information of two kinds: (a) information about causal dependencies and independencies between relevant events (corresponding to structural equations) and (b) information about some background facts (corresponding to the values of exogenous variables). For illustration, this is how the equations in the execution model get transposed into premises:[16]

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

Notice that every question in the pairs is related to the random variable appearing on the left-hand side of the equation. Also the information about exogenous variables is encoded in this form. This time the question in play has as its members the proposition itself and its negation. Assuming that the court does not issue the order, we get:

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

Setting up a premise semantics in this way involves a number of idealizations.[17] Here I don't investigate how to eliminate them: the reason is that I want to keep things simple on this end, and focus on different issues. But they can be relaxed; for example, some of the ideas in Kaufmann's (2013) semantics serve exactly this purpose.[18]

[16] For readability, I use italic letters to stand for the relevant propositions.

[17] Let me flag some of these idealizations. 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.

[18] Kaufmann's key maneuver is to impose a *structure* on premise sets. Rather than a set of individual propositions, he uses a set of sets of propositions. The relevant sets of propositions have to be closed under causal ancestors. For example, take the simple causal model depicted below:

Captain's order $\longrightarrow$ **X firing** $\longrightarrow$ **Prisoner's death**

Suppose that the equations of the model are

4.4 Basic filtering

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 us first introduce two definitions:

> **A proposition p is an answer to a premise** P iff $P = \langle Q, r\rangle$ and $p \in Q$.
>
> **A proposition p settles a premise** P iff, for some answer q to P, $p \vDash q$

With these definitions in hand, we can define the filtering of a premise set:

> A **filtering of a premise set Π relative to proposition** p (formally: '$\Pi|p$') is a premise set Π' such that, :
>
> (i) $\langle\{p, \bar{p}\}, p\rangle \in \Pi'$;
>
> (ii) for all premises $P \in \Pi$, if p doesn't settle P, $P \in \Pi'$;
>
> (iii) no other premises are in Π'.

In short, we build a filtered premise set Π' from an original premise set Π by (a) letting in the premise corresponding to the conditional antecedent, i.e. $\langle\{p, \bar{p}\}, p\rangle$; (b) carrying over any premise that is not settled by p.

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 Π, or Prop_Π. Formally:

> The **proposition set** of a premise set Π is the set Prop_Π 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 first-pass statement in (9)):

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

More informally: ⌜if ϕ, would ψ⌝ is true iff the premise set $g(w)$, filtered for ϕ, includes propositions that jointly entail ψ. Let me show how this works for a basic example. Consider again (7), repeated below:

X=C

D=X

and that C=X=D=1. Then the relevant premise set is, on Kaufmann's semantics

$\{\{c\}\{x, c\}\{d, x, c\}\}$

(where the lowercase variable stand for the positive values of the relevant random variables).
So far as I can see, my strategy of appealing directly to the values of the background variables (which closely mirrors Pearl's own procedure for evaluating counterfactuals) obtains equivalent results to Kaufmann's semantics, as long as (a) we consider only deterministic processes, and (b) the information contained in the equations is built in full into the ordering source.

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

Here is how the semantics handles (7). The initial premise set (on the left) gives rise to the premise set filtered for the antecedent (on the right, changes in boldface):

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

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

4.5 Summary

Filtering semantics differs from classical comparative closeness semantics in one key respect. On classical semantics, we evaluate a counterfactual by adding the antecedent to our stock of information, and we check all ways of making that stock consistent. On filtering semantics, we also remove some information from our existing stock. In a slogan, classical semantics employs a 'global' strategy for solving inconsistency ("check *all* ways to make the premise set consistent"), filtering semantics a 'local' strategy ("check *some* ways to make the premise set consistent, specifically the ones that ignore information about the causal links upstream from the antecedent").

This concludes my outline of the basic implementation of filtering semantics. The next section is devoted to refining the semantics, in the light of pretty obvious problems of this first-pass version. Readers not interested in these details may skip ahead to §6, where I discuss the empirical upshot of filtering semantics.

5 Filtering semantics: complications

5.1 Minimally different models

The basic semantics of §4 won't work. The reasons is that the filtering operation won't, in general, yield a unique result. There may be multiple ways to filter a set of premises for an antecedent. To see this, consider once more the prisoner scenario and take the counterfactual:

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

The antecedent of (10) doesn't trigger any filtering. Recall the premise set I've been using:

$$\begin{array}{ll} & \langle\{c, \bar{c}\}, c \leftrightarrow u\rangle \\ & \langle\{x, \bar{x}\}, x \leftrightarrow c\rangle \\ (11) & \langle\{y, \bar{y}\}, y \leftrightarrow c\rangle \\ & \langle\{d, \bar{d}\}, d \leftrightarrow (x \vee y)\rangle \\ & \langle\{u, \bar{u}\}, \bar{u}\rangle \end{array}$$

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 is not the result we want.[19]

Let me restate the problem informally. The key idea behind filtering is that we modify the background information that we use to evaluate a conditional. Our first-pass attempt simply assumes that each conditional antecedent settles how this information should be modified. This is too simplistic. Conditional antecedents may be too unspecific to determine exactly how the relevant information changes. The natural suggestion is that we consider multiple ways of modifying the background information in the light of the antecedent. In technical terms, the suggestion is that the semantics should consider *multiple ways of filtering the premise set*. This basic suggestion is simple enough, but we need some work to establish exactly what counts as an appropriate way to perform the filtering.

On a first pass, we might consider *all* ways to filter the premise set that make the conditional antecedent true.[20] In formal terms, this amounts to considering all ways of settling the questions in the premise set that entail the antecedent of the conditional. For example, if we use the antecedent of (10) to filter the premise set in (11), we proceed by considering all ways of settling the questions in (11) that make either x or y true. On the resulting proposal, a counterfactual is true iff the consequent is entailed by all filterings generated in this way.

It's easy to see that this is too strong. Consider the following set of answers:

$$\{u, c, x, y, \bar{d}\}$$

This way of settling the questions in the premise set makes x and y true; it also makes d false. This is a scenario where the court issues the execution order, the captain sends the signal to the riflemen, the riflemen shoot, and yet the prisoner doesn't die. If this scenario is relevant for evaluating (10), then (10) is false. But of course, intuitively, this is not a scenario we should consider when evaluating (10): the reason is that it violates our knowledge about the causal structure of the prisoner situation. We know that, given the causal setup of the situation, one of the two riflemen shooting is causally sufficient to kill the prisoner (keeping other assumptions unaltered). Hence a scenario where the riflemen shoot but the prisoner doesn't die should be irrelevant. Somehow or other, when evaluating (10) we should restrict consideration to other ways of settling answers to questions in the premise set.

The natural suggestion is that build on the 'minimal difference' intuition that is at the basis of counterfactual semantics since Stalnaker and Lewis. In the context of a causal models-based semantics, the 'minimal difference' intuition is naturally cashed out in terms of minimal antecedent-verifying interventions: i.e., interventions that change as few premises as possible, while still generating a premise set that entails the antecedent. Informally, and on a first pass, here are the new truth conditions of a counterfactuals:

> $\ulcorner$if ϕ, would $\psi\urcorner$ is true relative to a premise set Π iff all the minimal filterings of Π for ϕ also make ψ true.

[19] 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 (10) trivially true (or, if we tried to enforce a kind of nonvacuousness presupposition, defective).

[20] Thanks to an anonymous referee for pushing me to discuss explicitly this case.

As an aside, let me notice that this semantics resembles conditional semantics that make use of truthmakers[21] (while of course avoiding any appeal to a notion of metaphysical truthmaker). The role of truthmakers is played by sets of answers to the questions in the premise set: a counterfactual is true iff the consequent is true on all minimal ways of intervening on the answers in the question set that make the antecedent true.

The new version of the semantics includes an algorithm that implements these intuitive ideas in a formal system. For illustration, let me anticipate the result of the proposal for (10). The semantics considers the following two filterings—one for each of the disjuncts:

$$
\begin{array}{lcl}
 & & \langle \{c, \bar{c}\}, c \leftrightarrow u \rangle \\
 & & \langle \{\boldsymbol{x}, \bar{\boldsymbol{x}}\}, \boldsymbol{x} \rangle \\
 & \Longrightarrow & \langle \{y, \bar{y}\}, y \leftrightarrow c \rangle \\
\langle \{c, \bar{c}\}, c \leftrightarrow u \rangle & & \langle \{d, \bar{d}\}, d \leftrightarrow (x \vee y) \rangle \\
\langle \{x, \bar{x}\}, x \leftrightarrow c \rangle & & \langle \{u, \bar{u}\}, \bar{u} \rangle \\
\langle \{y, \bar{y}\}, y \leftrightarrow c \rangle & & \\
\langle \{d, \bar{d}\}, d \leftrightarrow (x \vee y) \rangle & & \langle \{c, \bar{c}\}, c \leftrightarrow u \rangle \\
\langle \{u, \bar{u}\}, \bar{u} \rangle & & \langle \{x, \bar{x}\}, x \leftrightarrow c \rangle \\
 & \Longrightarrow & \langle \{\boldsymbol{y}, \bar{\boldsymbol{y}}\}, \boldsymbol{y} \rangle \\
 & & \langle \{d, \bar{d}\}, d \leftrightarrow (x \vee y) \rangle \\
 & & \langle \{u, \bar{u}\}, \bar{u} \rangle
\end{array}
\tag{12}
$$

I call the premise sets resulting from this procedure **permissible filterings** of the original premise sets. Hence the new schematic truth conditions of a counterfactual are:

(13) $[\![\text{if } \phi, \text{ would } \psi]\!]^{w,g} = 1$ iff for every Π s.t. Π is a permissible filtering of $g(w)$ for ϕ, the proposition set of Π entails that ψ is true.

The rest of this section is dedicated to giving a precise formal definition of a permissible filtering and a formal statement of the semantics.

5.2 Permissible filterings

Before starting, one *caveat*. So far, the semantics I have developed stays close to the machinery shared by all versions of the causal models framework. At this point, I have to go beyond that. The basic version of the causal models framework can handle only a simple array of counterfactuals; for example, the theory doesn't settle how to handle counterfactuals whose antecedent doesn't exactly coincide with an answer to one of the relevant questions, like (10). I have chosen to develop a technology that is both intuitive and relatively conservative with respect to premise semantics. But it's only one of the available options.[22]

[21] For some recent versions of truthmaker semantics for conditionals, see Fine 2012a and 2012b, Yablo 2014. Among other things, filtering semantics goes *some* of the way towards vindicating the logic of Fine's semantics, though not all of the way.

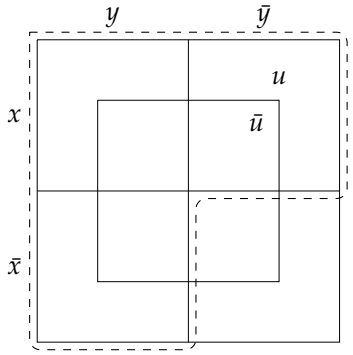
[22] Let me highlight two features of the semantics that are the result of taking choice points and that could be easily altered without affecting the basic idea. First, a disjunctive antecedent might induce more than two permissible filterings (in particular, we may add a permissible filtering that filters for both disjuncts); second, the quantification over permissible filtering may not be universal.

Here is the basic idea. We use something like the converse of the filtering algorithm we had in §4. There we checked whether the antecedent of a conditional settled the answer to any questions in the premise set. Now we check which answers or combinations of answers in the premise set entail the antecedent. In particular, we check which *minimal* combinations of answers (for some suitable way of understanding minimality) will make the antecedent true. This will capture the idea that filterings are minimal ways of modifying the premise set that make the antecedent true.

Before developing the formalism, let me illustrate the idea intuitively with diagrams. We start by considering the set of all the questions present in a premise set (call it 'question set'). For example, in the prisoner scenario, we get the following set:

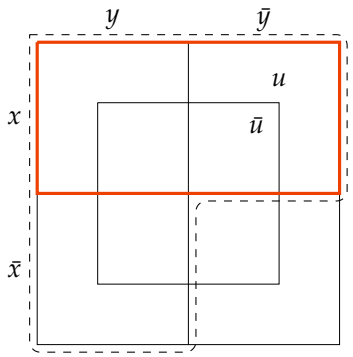
$$\Sigma_{\Pi} = \{\{u, \bar{u}\}, \{c, \bar{c}\}, \{x, \bar{x}\}, \{y, \bar{y}\}, \{d, \bar{d}\}\}$$

We can draw a diagram representing visually the questions in Σ_{Π}. What we get is a partition, where each cell represents a combination of answers. The first step for determining permissible filterings is locating the antecedent $\ulcorner x \vee y \urcorner$ on this partition via a dotted line. (I use only three questions to reduce visual clutter).

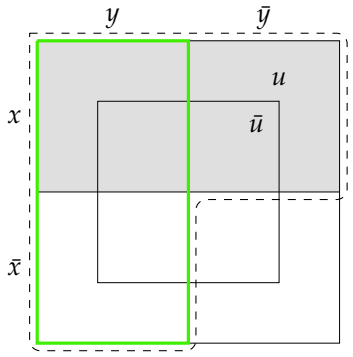


Then we check which cells in the diagram are wholly contained in the dotted area. This will tell us which cells make the antecedent true. In doing this, though, we don't consider immediately any old cell. We start by taking cells to be invisible. We make cells visible progressively by making more and more answers visible. As soon as we hit on a cell that is completely included in the dotted area, we mark the corresponding area with shading, write down the answers we made visible in the process, and then start afresh.

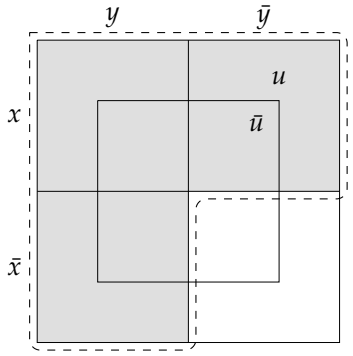
For example: suppose that we start by making visible the cell corresponding to answer x. (I mark visibility by using colored, thick borders.)



We observe that this cell is already fully contained in the dotted area. At this point, we make it invisible again and mark the corresponding area via the shading. We then proceed to make visible further cells. In our case, we will make y visible:



We stop when the whole dotted area is shaded. In our example, these two rounds are enough to cover the whole area.



$\{x\}$ and $\{y\}$ are the sets of answers we've used to cover the whole dotted area. It's easy to check that they are the smallest sets of answers we can use to do this. The various sets of cells that we've made visible in our two rounds will give us the permissible filterings. The set of these sets I call the *filter set* of the premise set, relative to the antecedent.

This should be enough to give an intuitive picture of how filtering works; I include more examples in a footnote.[23] Notice that, despite the heavy reliance on partitions, this filtering procedure still yields an intensional semantics: necessarily equivalent antecedents (and hence, among other things, logically equivalent antecedents) give rise to the same filterings, keeping fixed the premise set.

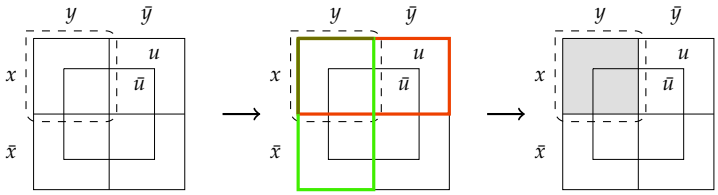
5.3 Formal semantics

Below I give the final version of filtering semantics in full detail. The exposition is technical, but I've already stated the main ideas informally via diagrams. So readers not interested in the formalism may skip ahead.

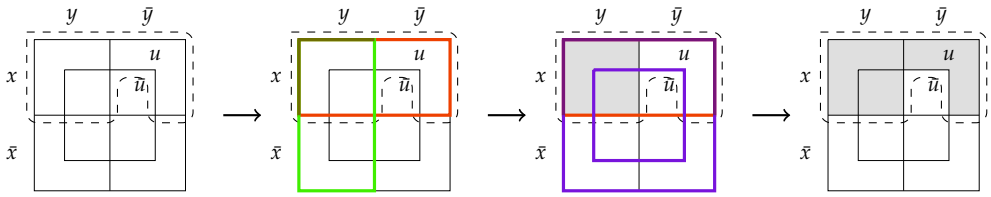
I start from a basic version of Kratzer semantics for modality. I assume an intensional system: the interpretation function is relativized to a world and an ordering source. Differently from Kratzer, I don't need a modal base: the function of the modal base in Kratzer's algorithm is subsumed in the functioning of the filtering algorithm.[24]

The semantics is static, hence denotations of clauses are standard propositions, taken to be functions from worlds to truth values. This assumption serves just the purposes of simplicity. So far as I can see, all relevant changes could be exported in full into a dynamic system, in which counterfactuals are taken to denote functions for updating the context (e.g. in the style of von Fintel 2001 and Gillies 2007).[25]

[23] Consider first the conjunctive antecedent $\ulcorner x \wedge y \urcorner$. In this case, we need to make visible both x and y together to get a cell that is enclosed in the dotted area. Once we do this, the whole dotted area is covered.



Hence the antecedent $\ulcorner x \wedge y \urcorner$ generates only one permissible filtering, i.e. the one resulting from filtering both $\langle \{x, \bar{x}\}, x \leftrightarrow c \rangle$ and $\langle \{y, \bar{y}\}, y \leftrightarrow c \rangle$.
Finally, let me illustrate a more complex Boolean compound, i.e. $\ulcorner x \wedge (y \vee u) \urcorner$. In this case, we have to make visible two answers at a time to get a cell that is included in the dotted area. In the diagram below, I first make visible x and y, and then x and u.



Hence the antecedent $\ulcorner x \wedge (y \vee u) \urcorner$ triggers two permissible filterings. On the first, we filter out both $\langle \{x, \bar{x}\}, x \leftrightarrow c \rangle$ and $\langle \{y, \bar{y}\}, y \leftrightarrow c \rangle$, on the second both $\langle \{x, \bar{x}\}, x \leftrightarrow c \rangle$ and $\langle \{u, \bar{u}\}, \bar{u} \rangle$. Notice that these are the same filterings we would get from the equivalent $\ulcorner (x \wedge y) \vee (x \wedge u) \urcorner$ proposition.

[24] Notice: this will change if we give up Kratzer's claim (in 1981b) that the modal base starts empty for counterfactuals. It is a straightforward task to generalize filtering semantics to that case; I leave that as an exercise to the reader.

[25] As in §2, for simplicity I allow myself to be sloppy and use the labels 'modal base' and 'ordering source' both for f

Before getting to the semantics proper, I must lead you through a few definitions. I already used the notion of a *question set*: the *question set* of a premise set Π is simply the set of all questions appearing in the premise set:

(14) $\Sigma_\Pi = \{Q : \exists P \in \Pi : \text{for some } p, P = \langle Q, p\rangle\}$

It's useful to define the answer set of a question set. The answer set is just the set of all the answers appearing in the question set. Since question sets are sets of sets of propositions, formally the answer set is just the union set of its question set.

(15) $\mathrm{A}_\Pi = \bigcup \Sigma_\Pi$

Notice that the question set is a set of sets of propositions, while the answer set is just a set of propositions.

Here is how I capture filtering. Given a counterfactual antecedent p, we single out the *minimal subsets of the answer set* A_Π that entail p. The set of these subsets I call the filter set of a premise set, relative to an antecedent.

> The **filter set** of a premise set Π relative to proposition p is the set $\Phi_{\Pi,p}$ of all minimal subsets S' of the answer set A_Π such that $S' \vDash p$

In symbols:

(16) $\Phi_{\Pi,p} = \{S' \subseteq \mathrm{A}_\Pi : S' \vDash p \text{ and } \neg\exists S'' : S'' \subset S' \text{ and } S'' \vDash p\}$

Finally, we define the notion of a permissible filtering. Informally, a permissible filtering of a premise set Π relative to a proposition p is the result of (a) picking a set member of the filter set and (b) filtering out all and only the premises whose questions are answered by that set of propositions, while letting in the premise corresponding to p.

Let us start working towards a precise definition. First, recall our definition of a proposition answering a premise, from §4.4.

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

Now we're ready to define permissible filterings:

> A **permissible filtering** of a premise set Π relative to proposition p is a premise set Π_p such that:
>
> (i) $\langle\{p, \bar{p}\}, p\rangle \in \Pi_p$;
>
> (ii) for some set of propositions S in the filter set $\Phi_{\Pi,p}$ and for all $P \in \Pi$:
>
> - if P is not answered by any proposition in S, $P \in \Pi_p$;
> - if P is answered by some q in S, $\langle\{q, \bar{q}\}, q\rangle \in \Pi_p$;
>
> (iii) nothing else is in Π_p.

and g proper, and for the sets $f(w)$ and $g(w)$, determined by plugging the world of evaluation w as an argument of the two functions. I trust that the reader will be able to disambiguate when needed.

To summarize: we first let in the premise corresponding to the proposition we filter for; then, if a premise is not filtered out, it is carried over from the premise set to the permissible filtering; if it is filtered out, it is replaced by a premise whose proposition is just the relevant answer.

It's useful to go through an example in detail. Take again the counterfactual:

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

Let the relevant ordering source be $g(w)$. Here is the question set of $g(w)$ (repeated from above):

(17) $\Sigma_{g(w)} = \{\{u, \bar{u}\}, \{c, \bar{c}\}, \{x, \bar{x}\}, \{y, \bar{y}\}, \{d, \bar{d}\}\}$

Here is the answer set:

(18) $\mathrm{A}_{g(w)} = \{u, \bar{u}, c, \bar{c}, x, \bar{x}, y, \bar{y}, d, \bar{d}\}$

We have already seen the filter set of $g(w)$ with respect to the antecedent of (10). It is the set containing the singletons of x and y:

(19) $\Phi_{g(w),x\vee y} = \{\{x\}, \{y\}\}$

Correspondingly, there are two permissible filterings of $g(w)$—i.e., those in (12)), which I repeat below.

(12)
$$\begin{array}{lcl} & & \langle\{c,\bar{c}\}, c \leftrightarrow u\rangle \\ & & \langle\{\boldsymbol{x},\bar{\boldsymbol{x}}\}, \boldsymbol{x}\rangle \\ & \Longrightarrow & \langle\{y,\bar{y}\}, y \leftrightarrow c\rangle \\ \langle\{c,\bar{c}\}, c \leftrightarrow u\rangle & & \langle\{d,\bar{d}\}, d \leftrightarrow (x \vee y)\rangle \\ \langle\{x,\bar{x}\}, x \leftrightarrow c\rangle & & \langle\{u,\bar{u}\}, \bar{u}\rangle \\ \langle\{y,\bar{y}\}, y \leftrightarrow c\rangle & & \\ \langle\{d,\bar{d}\}, d \leftrightarrow (x \vee y)\rangle & & \langle\{c,\bar{c}\}, c \leftrightarrow u\rangle \\ \langle\{u,\bar{u}\}, \bar{u}\rangle & & \langle\{x,\bar{x}\}, x \leftrightarrow c\rangle \\ & \Longrightarrow & \langle\{\boldsymbol{y},\bar{\boldsymbol{y}}\}, \boldsymbol{y}\rangle \\ & & \langle\{d,\bar{d}\}, d \leftrightarrow (x \vee y)\rangle \\ & & \langle\{u,\bar{u}\}, \bar{u}\rangle \end{array}$$

Now, finally, to the semantics proper.[26] Here is a schematic entry:

(20) $[\![\text{if } \phi, \text{ would } \psi]\!]^{w,g} = 1$ iff for every premise set $\Pi_{\|\phi\|_g}$ s.t. $\Pi_{\|\phi\|_g}$ is a permissible filtering of $g(w)$ relative to $\|\phi\|_g$, the proposition set of $\Pi_{\|\phi\|_g}$ entails $\|\psi\|_g$

As usual in premise semantics, this mechanism can be easily generalized to modals with different quantificational force. We just need to use logical relations different from entailment:

[26]For simplicity, I just give a syncategorematic meaning for the modal *would*. I ignore all issues concerning tense, including issues about the presence of the past tense in *would*. I am sympathetic to views on which *would* is not a semantic unit, but rather should be decomposed into a modal auxiliary and a past tense. (For examples of accounts based on this view, see, among many, Iatridou 2000, Condoravdi 2002, and Kaufmann 2005.) But I leave it to future research to integrate the current account with this view.

for example, following Kratzer, we can use compatibility to capture the meaning of *might*-counterfactuals. Here is a first-pass sample entry for *might*-counterfactuals.

(21) $[\![\text{if } \phi, \text{ might } \psi]\!]^{w,g} = 1$ iff for every premise set $\Pi_{\|\phi\|_g}$ s.t. $\Pi_{\|\phi\|_g}$ is a permissible filtering of $g(w)$ relative to $\|\phi\|_g$, the proposition set of $\Pi_{\|\phi\|_g}$ is compatible with $\|\psi\|_g$

6 Filtering semantics: empirical aspects

6.1 A new semantics?

I have spent §4 and §5 setting up filtering semantics. But how different is really filtering semantics from standard premise semantics? In particular, couldn't we simulate the old semantics somehow by means of the old semantics? Lewis (1979) refused to take a notion of causal dependence as primitive when specifying the ordering employed by the semantics for counterfactuals. But we might try, against Lewis, to simulate the functioning of filtering semantics by using explicitly causal information in the ordering source. One natural thought is that a causal models-based semantics will be properly matched by an old-style premise semantics with a causal ordering source.

But simulating filtering semantics is not so easy. There are real divergences between classical semantics and filtering semantics that cannot be just eliminated by cherry-picking the ordering or the premise set. Let me highlight two of them.

The first is that, as I'm about to show, we get a different logic. This won't change by exploiting a different ordering or premise set. A semantics fixes a logic via its structural features, like the quantifiers in play and the formal properties of the ordering relation. How we interpret the ordering relation doesn't matter. As a result, we will still have a substantial difference in empirical predictions. In fact, arguments from the logic have always been the strongest considerations at our disposal to decide between different semantics of counterfactuals. The success of comparative closeness semantics is due just to its capacity to account for logical properties of counterfactuals that competing views (such as, for example, a naïve strict conditional analysis) couldn't predict. Insofar as we have differences in the logic, we will have similar arguments one way or the other.

The second difference is more philosophical in nature and concerns the kind of information that is employed by a semantics. Standard semantics for counterfactuals employs intensional information to generate an ordering on worlds: to see this, notice that the information that is included by Kratzer's ordering source consists in ordinary possible worlds propositions. Filtering semantics requires more structure. The new ordering source employs directional premises: i.e., premises that include information about what determines what, according to some relevant determination relation.[27] Hence, while filtering semantics is still intensional (in the sense that

[27]Notice that, while throughout the paper I have stuck to the usual causal interpretation of Pearl's framework, nothing in the formal part of the theory dictates that we hold on to this interpretation. (In fact, philosophers have started to find applications for the causal models framework that go beyond the causal case—see, for example, Wilson 2013.) What really distinguishes the premises used in filtering semantics from those used in standard semantics is the presence of direction of determination, and not a specific reference to causation.

it yields the same truth and falsity verdicts for necessarily equivalent antecedents and consequents), it appeals to resources that go beyond standard intensional information. In particular, its use of partitions and questions to capture determination relations establishes an interesting link with a metaphysics that exploits grounding or other, related dependence relations. If filtering semantics turns out to be correct, then our counterfactual thought and talk will turn out to involve something like a notion of dependence, at least in structure.

6.2 A problem

Let me illustrate one point in which the logics generated by filtering semantics and standard premise/ordering semantics diverge.[28] 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 three were invited. None of them ended up going, but each of them kept track of whether the person they liked was going. Each of them wanted an occasion to spend time with their beloved and without their suitor. Having an occasion of this kind would have been sufficient for each of them to go.

I claim that, on at least one reading, (22) is judged true, while (23) is judged false, or at least dubious.

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

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

By symmetry, we get the following set of judgments. On at least one reading, these counterfactuals (call them 'forward loop counterfactuals') are judged true:

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

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

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

On the same reading, these counterfactuals (call them 'backward loop counterfactuals') are judged oftentimes false, or dubious:

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

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

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

[28] As I mentioned in the introduction, the fact that a causal-models-based logic and standard counterfactual logics diverge has emerged very recently in the literature on causal models. Halpern 2013 shows that, as long as we extend consideration to a wide enough class of causal models (i.e. all causal models that, for all choices of exogenous variables, have a unique solution), the logics will differ. I learned of Halpern's result only after independently discovering the evidence that instances of Loop seem to fail natural language.

Let me emphasize that I only have in mind one possible reading of the counterfactuals (to me and other informants, this reading is the one that is naturally suggested by the question 'What would happen if Andy/Billy/Charlie was at the party?'). For all I need here, it may well be that all of (22)–(27) have different truth values on other readings. This doesn't affect my argument.

To summarize: if my empirical claim is correct, there is a reading on which we get the following configurations of judgments:

$$\begin{array}{ll} A \boxrightarrow B \checkmark & B \boxrightarrow A \text{ ✗} \\ C \boxrightarrow A \checkmark & A \boxrightarrow C \text{ ✗} \\ B \boxrightarrow C \checkmark & C \boxrightarrow B \text{ ✗} \end{array}$$

The problem is simple: it is impossible to accommodate these 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 strict 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 strict 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 $A \boxrightarrow B$, w^* is also a B-world. Since $B \boxrightarrow C$, 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, $A \boxrightarrow C$ is true. *QED.*

The proofs for other versions of ordering and premise semantics are more involved, but have the same structure. (The reader can consult Halpern's 2013 for the proof concerning Lewis-style semantics with limit assumption).

Before proceeding, let me quickly block a way of dismissing the data. The thought would be simply to invoke context shift: forward loop counterfactuals would be evaluated with respect to one ordering source; backward loop counterfactuals with respect to another. Obviously, on these assumptions, standard Kratzer semantics can yield the right predictions for all of (22)–(27). But this is not a good reply, for two reasons. First, appealing to context dependence without independent motivation, and when a systematic account is available, is bad methodology. Second, in this case in particular, not only do we lack reason to suspect a context change, but in fact we have reason to think that there isn't one.

As for the first point: Lewis himself pointed out in *Counterfactuals* that we can always appeal to context shifts to accommodate problematic data. By making enough assumptions about context shifts, we can even dismiss all the data motivating his original comparative closeness semantics. But this, as he puts is, "defeatist"; for

> it consigns to the wastebasket of contextually resolved vagueness something much more amenable to systematic analysis than most of the mess in that wastebasket (1973a, p. 13).

Lewis's point applies in full to our scenario. In the case of (22)–(27), we have no independent reason to think that there is a context shift. Indeed, it would seem extraordinary that context should systematically shift just when we evaluate backward loop counterfactuals. Why should speakers be naturally inclined to favor one reading for the case of forward loop counterfactuals, and a different reading for backward loop counterfactuals?

Second, it is a common assumption in semantics that, when a sentence has different readings that differ in truth value, *ceteris paribus* speakers tend to focus on a true reading. This entails that, if there is a single ordering source that makes all loop counterfactuals true, speakers will tend to default to it when giving judgments. Yet this doesn't happen. This is evidence that we have a semantic phenomenon that cannot be dealt simply with with context shifts.

6.3 Loop

Here is a more general way of stating the problem. Consider the following inference rule:

$$\textsc{Loop}\quad \frac{\begin{array}{l}\phi \mathrel{\Box\!\!\rightarrow} \psi \\ \psi \mathrel{\Box\!\!\rightarrow} \chi \\ \chi \mathrel{\Box\!\!\rightarrow} \phi\end{array}}{\phi \mathrel{\Box\!\!\rightarrow} \chi}$$

Loop is a valid rule in the logics generated by classical premise semantics, as well as in all standard counterfactual logics. In fact, something stronger is true: Loop is an instance of a general rule schema, which I call Generalized Loop. All rules that are instances of Generalized Loop are valid in classical counterfactual semantics.

$$\textsc{Generalized Loop}\quad \frac{\begin{array}{l}\phi_1 \mathrel{\Box\!\!\rightarrow} \phi_2 \\ \phi_2 \mathrel{\Box\!\!\rightarrow} \phi_3 \\ \ldots \\ \phi_{k-1} \mathrel{\Box\!\!\rightarrow} \phi_k \\ \phi_k \mathrel{\Box\!\!\rightarrow} \phi_1\end{array}}{\phi_1 \mathrel{\Box\!\!\rightarrow} \phi_k}$$

To my knowledge, Loop and Generalized Loop haven't been discussed in any detail in the literature on counterfactuals, either in philosophy or in semantics. They do appear in the literature on belief revision and nonmonotonic logic: see Kraus et al. 1990.[29] But Loop and Generalized Loop are important for my purposes because they show the point of divergence between filtering and classical premise semantics. While they are valid in standard premise semantics, they are invalid in filtering semantics.

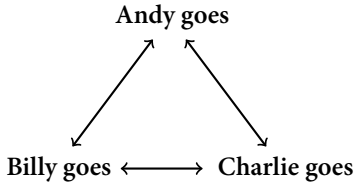
To see how filtering semantics invalidates Loop, start by considering a simple causal model for the party scenario:[30]

[29] Interestingly, in Kraus, Lehmann, and Magidor's models, the validity of Generalized Loop is just what forces the relation of comparative closeness in play in their models to be transitive. Given that transitivity is necessary for having an ordering, this seems to suggests that there is a connection between the validity of Generalized Loop and the semantics making use of an ordering. This is just a conjecture, though; I don't know of any general result that establishes this point.

[30] 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.

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. This can be seen very easily by looking at the graph, which is cyclic:

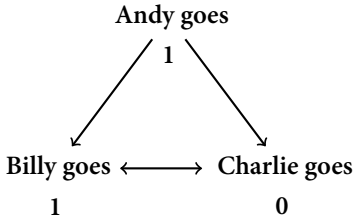


At the same time, the model does contain information that is sufficient to determine the values of the relevant variables: in particular, the model has a unique solution, the one on which all the variables have value 0. [31] In addition to this, and importantly for our purposes, the model yields exactly the intuitive verdicts when it is used to evaluate the relevant counterfactuals. To see this, consider how (22) (repeated below) is evaluated:

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

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

A = 1
B = (A $\wedge\neg$C)
C = (B $\wedge\neg$A)



As the graph shows, in the modified model B must have value 1 and C value 0. We will get analogous results, *mutatis mutandis*, by intervening on each of the relevant variables in the model.

It's easy to see that we get analogous results on filtering semantics. Forward loop counterfactuals ((22), (24), and (25)) are predicted to be true; backwards loop counterfactuals ((26), (23), and (27)) are predicted to be false. I illustrate in detail the case of (22) and (26); other cases are perfectly symmetrical.

This is the initial premise set:

(28) $\langle\{a,\bar{a}\}, a \leftrightarrow (c \wedge \neg b)\rangle$
$\langle\{b,\bar{b}\}, b \leftrightarrow (a \wedge \neg c)\rangle$
$\langle\{c,\bar{c}\}, c \leftrightarrow (b \wedge \neg a)\rangle$

[31] 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.

The antecedent of both (22) and (26) generates only one permissible filtering, namely:

(29) $\langle \{\boldsymbol{a}, \bar{\boldsymbol{a}}\}, \boldsymbol{a} \rangle$
$\langle \{b, \bar{b}\}, b \leftrightarrow (a \wedge \neg c) \rangle$
$\langle \{c, \bar{c}\}, c \leftrightarrow (b \wedge \neg a) \rangle$

It's easy to check that the propositions in the premises in (29) entail b and $\bar{c}$, thus yielding the intuitively right predictions for (22) and (26).

Notice that filtering is crucial to get this result. If we evaluated the counterfactuals by using the same ordering source (*modulo* the difference in the type of the premises), but using classical premise semantics, all counterfactuals (22)–(27) would be predicted to be false, and all the corresponding *might* counterfactuals would be predicted to be true.

6.4 Remarks on filtering logic

The premise set I give above witnesses the failure of Loop on filtering semantics. Since Loop is valid on standard ordering/premise semantics, this shows that filtering semantics gives rise to a different logic. This is not the place to pursue a detailed study of the logic generated by the new semantics. But let me point out where the new logic departs from standard counterfactual logic. For ease of reference, I will use the labels of one of the standard axiomatizations of counterfactual logic, namely the one in Burgess 1981.

Given the semantics that I have developed in this paper, the element to reject is Burgess's axiom (A4):

(A4) $((\phi \boxrightarrow \chi) \wedge (\psi \boxrightarrow \chi)) \supset ((\phi \vee \psi) \boxrightarrow \chi)$

Kraus et al. 1990 point out that (A4), together with some very basic assumptions, allows the derivation of Loop. Hence it's unsurprising that the axiom is invalid in the new semantics. To see how filtering semantics generates a counterexample to (A4), take again the prisoner scenario, and consider the following three counterfactuals:

(25) If Billy was at the party, Charlie would be at the party.

(30) If Andy was at the party and Billy wasn't at the party, Billy would not be at the party.

(31) If Andy was at the party or Billy was at the party, either Billy would not be at the party or Charlie would be at the party.

(25) is true, given our choice of premises. (30), which is of the form $\ulcorner (A \wedge \neg B) \boxrightarrow \neg B \urcorner$ is true on any choice of premises on any plausible counterfactual semantics. (31), on the other hand, comes out false on the premises given above. Given these assignments of truth value, we can use simple inference rules to generate a counterexample to (A4) (details are in a footnote).[32]

[32] In its current setup, the semantics validates the rule of substitution of provable equivalents mentioned just below in the main text, as well as the following rule (I borrow the name from Fine 2012a):

(right weakening) $(\phi \boxrightarrow (\psi \wedge \chi)) \supset (\phi \boxrightarrow \chi)$

By (right weakening) and the substitution rule used on the consequents of the conditionals we get, from (25) and (30), respectively:

I should flag that there is another option, though a more far-fetched on from the current standpoint. This option involves restricting replacement of provable equivalents. The following rule holds in Burgess's system:

(RPE) From $\gamma \leftrightarrow \delta$ and β infer α, where $\alpha = \psi(\delta)$ differs from $\beta = \psi(\gamma)$ only by replacing some subformulas of β of form γ by δ.

We might block the derivability of Loop by rejecting the validity of substitution of provable equivalents in the antecedents of counterfactuals. (This route bears interesting similarities to a recent proposal put forward by Kit Fine in the context of a very different framework; see his 2012a, 2012b.) But this would involve a greater departure from classical systems, and a switch to a fully hyperintensional semantics.[33]

7 Conclusion

The goal of this paper has been to show how causal-models-inspired ideas can be implemented in a possible worlds semantics for counterfactuals. I have focused specifically on one aspect of this implementation—namely, the algorithm for resolving inconsistency that is at work in causal models. I have showed that implementing this algorithm yields a new kind of possible worlds semantics, which generates a new logic.

Let me close by mentioning two issues that I haven't covered here. The first, which I briefly hinted at in §2.4, has to do with counterfactuals that track noncausal connections. The current version of filtering semantics uses information about causal dependencies and independencies. Hence it's unclear how it would handle noncausal counterfactuals. One natural thought is that structural equations models are general enough to capture all sorts of dependence relations. Some philosophers have put forward arguments to this effect: see, for example, Wilson 2013 and Schaffer 2015. Similarly, a premise semantics that exploits directional premises may be used to model also noncausal dependence. Of course, the details of this implementation remain to be worked out.

(i) $b \boxrightarrow (\neg b \vee c)$

(ii) $(a \wedge \neg b) \boxrightarrow (\neg b \vee c)$

Similarly, from (31) we can get, via substitution (in the antecedent, this time):

(iii) $((a \wedge \neg b) \vee b) \boxrightarrow (\neg b \vee c)$

It's easy to see that the fact that (i) and (ii) are predicted to be true and (iii) is predicted to be false is incompatible with an instance of (A4) (just insert b for ϕ, $(a \wedge \neg b)$ for ψ, and $(\neg b \vee c)$ for χ).

[33] The point is connected to a further feature of the logic, namely the validation of the inference rule that Fine calls 'simplification':

$$\frac{\phi \vee \psi \boxrightarrow \chi}{\phi \boxrightarrow \chi,\ \psi \boxrightarrow \chi}$$

Simplification is not validated by standard counterfactual logics. Yet it was pointed out early on in responses to Lewis (Fine 1975, Nute 1975), that it is an intuitively valid principle in natural language. Interestingly, filtering semantics goes close to validating simplification: for a special case in which the inference holds, consider just the example of a disjunctive antecedent in §5. But it doesn't quite vindicate Simplification across the board.

The second issue has to do with so-called backtracking counterfactuals. With some approximation, backtracking counterfactuals are counterfactuals that involve an epistemic inference between the antecedent and the consequent. An example of a backtracker, with reference to the prisoner scenario, would be:

(32) If the prisoner had died, one of the two riflemen (or both) would have shot.

(32) has a true reading (presumably in addition to a false one). Causal models, at least in their basic version, are notoriously unable to capture this reading; similarly for the version of filtering semantics that I have presented in this paper. Also in this case, there is a natural idea we can appeal to: backtracking readings might be the ones where the point of intervention is shifted 'upstream'.[34] In other words, the evaluation of counterfactuals works still in the usual way—we remove contradiction-generating information from our stock. Only, it is not information about dependencies that are immediately upstream from the antecedent, but information about dependencies higher up. Also in this case, a proper implementation must wait for a different occasion.

[34]For an interventionist-friendly view in this vein, see Dehghani et al. 2012). For studies that criticize the interventionist framework just because of its difficulties with backtracking counterfactuals, see, among many, Rips 2010.

References

Briggs, Rachael (2012). "Interventionist Counterfactuals." *Philosophical studies*, 160(1): pp. 139–166.

Burgess, John P (1981). "Quick Completeness Proofs for Some Logics of Conditionals." *Notre Dame Journal of Formal Logic*, 22(1): pp. 76–84.

Cariani, Fabrizio, Magdalena Kaufmann, and Stefan Kaufmann (2013). "Deliberative Modality under Epistemic Uncertainty." *Linguistics and Philosophy*, 36(3): pp. 225–259.

Chisholm, Roderick M (1946). "The Contrary-to-Fact Conditional." *Mind*, 55(219): pp. 289–307.

Condoravdi, Cleo (2002). "Temporal Interpretation of Modals: Modals for the Present and for the Past." In D. Beaver, S. Kaufmann, B. Clark, and L. Casillas (eds.) *The Construction of Meaning*, Palo Alto, CA: CSLI Publications.

Dehghani, Morteza, Rumen Iliev, and Stefan Kaufmann (2012). "Causal explanation and fact mutability in counterfactual reasoning." *Mind & Language*, 27(1): pp. 55–85.

Fine, Kit (1975). "Review of Lewis' Counterfactuals." *Mind*, 84: pp. 451–458.

Fine, Kit (2012a). "Counterfactuals Without Possible Worlds." *Journal of Philosophy*, 109(3): pp. 221–246.

Fine, Kit (2012b). "A Difficulty for the Possible Worlds Analysis of Counterfactuals." *Synthese*, 189(1): pp. 29–57.

von Fintel, Kai (2001). "Counterfactuals in a Dynamic Context." *Current Studies in Linguistics Series*, 36: pp. 123–152.

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

Gillies, Anthony S (2007). "Counterfactual scorekeeping." *Linguistics and Philosophy*, 30(3): pp. 329–360.

Goodman, Nelson (1947). "The Problem of Counterfactual Conditionals." *The Journal of Philosophy*, 44(5): pp. 113–128.

Groenendijk, Jerome, and Martin Stokhof (1984). *Studies in the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.

Halpern, Joseph (2000). "Axiomatizing Causal Reasoning." *Journal of Artificial Intelligence Research*, 12: pp. 317–337.

Halpern, Joseph Y. (2013). "From Causal Models to Counterfactual Structures." *Review of Symbolic Logic*, 6(2): pp. 305–322.

Hiddleston, Eric (2005). "A Causal Theory of Counterfactuals." *Noûs*, 39(4): pp. 632–657.

Iatridou, Sabine (2000). "The Grammatical Ingredients of Counterfactuality." *Linguistic Inquiry*, 31(2): pp. 231–270.

Kaufmann, Stefan (2005). "Conditional Truth and Future Reference." *Journal of Semantics*, 22(3): pp. 231–280.

Kaufmann, Stefan (2013). "Causal Premise Semantics." *Cognitive science*, 37(6): pp. 1136–1170.

Kratzer, Angelika (1981a). "The Notional Category of Modality." In H. J. Eikmeyer, and H. 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): pp. 201–216.

Kratzer, Angelika (1986). "Conditionals." In *Chicago Linguistics Society: Papers from the Parasession on Pragmatics and Grammatical Theory*, vol. 22, pp. 1–15. University of Chicago, Chicago IL: Chicago Linguistic Society.

Kratzer, Angelika (1989). "An Investigation of the Lumps of Thought." *Linguistics and Philosophy*, 12(5): pp. 607–653.

Kratzer, Angelika (1991). "Modality." *Semantics: An international handbook of contemporary research*, pp. 639–650.

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

Kraus, Sarit, Daniel Lehmann, and Menachem Magidor (1990). "Nonmonotonic Reasoning, Preferential Models and Cumulative Logics." *Artificial intelligence*, 44(1): pp. 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): pp. 418–446.

Lewis, David K. (1979). "Counterfactual dependence and time's arrow." *Noûs*, 13(4): pp. 455–476.

Lewis, David K. (1981). "Ordering Semantics and Premise Semantics for Counterfactuals." *Journal of Philosophical Logic*, 10(2): pp. 217–234.

Lewis, David K. (1982). "Logic for Equivocators." *Noûs*, 16(3): pp. 431–441.

Lewis, David K. (1988). "Relevant Implication." *Theoria*, 54(3): pp. 161–174.

Nute, Donald (1975). "Counterfactuals and the Similarity of Words." *The Journal of Philosophy*, 72(21): pp. 773–778.

Pearl, Judea (2000). *Causality: models, reasoning and inference*. Cambridge University Press.

Pollock, John L (1976). *Subjunctive Reasoning*. D. Reidel.

Rips, Lance J (2010). "Two causal theories of counterfactual conditionals." *Cognitive science*, 34(2): pp. 175–221.

Santorio, Paolo (2016). "Decentering Conditionals: Probability, Modus Ponens, and Stalnaker's Thesis." Manuscript, University of Leeds.

Schaffer, Jonathan (2015). "Grounding in the Image of Causation." *Philosophical Studies*.

Schulz, Katrin (2011). "If You'd Wiggled A, then B Would've Changed." *Synthese*, 179(2): pp. 239–251.

Slote, Michael A (1978). "Time in Counterfactuals." *The Philosophical Review*, 87(1): pp. 3–27.

Stalnaker, Robert (1968). "A Theory of Conditionals." In N. Recher (ed.) *Studies in Logical Theory*, Oxford.

Veltman, Frank (1976). "Prejudices, Presuppositions, and the Theory of Counterfactuals." In *Amsterdam Papers in Formal Grammar. Proceedings of the 1st Amsterdam Colloquium*, pp. 248–281. University of Amsterdam.

Wilson, Alastair (2013). "Metaphysical Causation." Manuscript, University of Birmingham.

Yablo, Stephen (2014). *Aboutness*. Princeton, NJ: Princeton University Press.