Causal Models and the Logic of Counterfactuals*

Jonathan Vandenburgh
Northwestern University

Abstract

Causal models provide a framework for making counterfactual predictions, making them useful for evaluating the truth conditions of counterfactual sentences. However, current causal models for counterfactual semantics face logical limitations compared to the alternative similarity-based approaches: they only apply to a limited subset of counterfactuals and the connection to counterfactual logic is not straightforward. This paper offers a causal model for the semantics of counterfactuals which improves upon these logical issues. It extends the causal approach to counterfactuals to handle more complex counterfactuals, including backtracking counterfactuals and those with logically complex antecedents. It also uses the notion of causal worlds to define a selection function and shows that this selection function satisfies familiar logical properties. While some limitations still arise, especially regarding counterfactuals which require breaking the laws of the causal model, this model improves upon many of the existing logical limitations of causal models.

Counterfactual conditionals like 'If A were the case, then C would be the case', written $A > C$, have two conflicting models for their truth conditions. Similarity-based models, like those of Lewis (2013) and Stalnaker (1968), assume that we have a set of possible worlds W with a similarity relation on worlds, $\leq_w$. They propose that the counterfactual $A > C$ is true at world w if C is true in all of the worlds closest to w (according to $\leq_w$) where A is true. Causal models, following Galles and Pearl (1998) and Pearl (2009), assume that we have a set of variables V with a causal model $\mathcal{M}$ describing how the variables in V are related to each other; a counterfactual $A > C$ is true if intervening on the causal description of the world to force A to be true entails that C is true.

One of the main advantages of causal models is that the models are determinate and cognitively realistic. Causal modeling builds on methods of statistical inference prevalent in epidemiology and econometrics, and elements of causal models (such as variable identification, structural equations, and residual or error terms) are frequently found in empirical work on counterfactuals and causal inference. Economists, for example, use elements of causal modeling to make counterfactual predictions for what would have happened if certain countries did not join the EU (Campos et al., 2019) or if video game companies had not

*Comments are welcome at jonathanvandenburgh2021@u.northwestern.edu.

developed games exclusively compatible with one platform (Lee, 2013). Furthermore, extensive psychological evidence supports the use of causal models to study human reasoning.[1] In contrast, similarity-based models are often too intractable to play the same role in empirical research or psychological theory.

However, philosophers have often preferred similarity-based approaches to causal models because they are more general, offering predictions for a broader range of counterfactuals, and because they correspond nicely to counterfactual logics. While similarity-based models can handle all kinds of counterfactual antecedents, causal models are usually limited in the antecedents they offer predictions for. The original causal model of Pearl, for example, could only handle counterfactuals with antecedents which are conjunctions of variable assignments. Furthermore, Pearl's model struggles with backtracking counterfactuals, where the antecedent is the effect rather than the cause of the consequent ('If the grass were wet, then it must have rained'). Hiddleston (2005) offers a framework which works for backtracking counterfactuals, but it struggles to explain some forward counterfactuals and still only applies to antecedents which are conjunctions of variable assignments. All causal models, furthermore, have difficulties explaining counterlegal counterfactuals, where the antecedent of the counterfactual breaks causal laws in the causal model.

The connection between these causal models and counterfactual logic is also underexplored. Pearl proves that his framework is sound and complete with respect to Lewis's axiomatization of counterfactual logic, **VC**, but this only holds for antecedents which are conjunctions of variable assignments.[2] Briggs (2012) offers an extension of Pearl's framework to more complex antecedents, but this logic differs from that of most counterfactual logics; modus ponens, for example, does not hold. Hiddleston's theory, like Pearl's theory, is restricted to antecedents which are conjunctions of variable assignments, and the connection between his framework and counterfactual logic is not developed in any detail.

In this paper, I offer a causal model for counterfactuals which applies to a wide class of counterfactuals, including those with antecedents of arbitrary logical complexity and backtracking counterfactuals, and develop the connection with similarity-based approaches and counterfactual logic. In §1, I introduce the foundations of causal models, following the interventionist approach of Pearl. I focus particularly on the distinction between exogenous and endogenous variables, an element of Pearl's framework which is often overlooked in philosophical discussions, and the need to incorporate backtracking counterfactuals. In §2, I develop an exogenous intervention model for counterfactuals which applies to antecedents of arbitrary logical complexity. I make use of the notion of a causal world and formally define a set of interventions which associates a world u and antecedent A with a set of intervened worlds which set A true. This set of intervened worlds serves as the selection function for the counterfactual semantics. In §3, I show that this selection function yields a familiar logic of

[1] See, for example, Glymour (2001), Sloman (2005), and Gopnik and Schulz (2007).

[2] Halpern (2000, 2013) argues that Pearl's proof is flawed because it ignores relevant axioms from Lewis's logic involving disjunctions. However, he provides a different proof that the result holds for recursive models.

counterfactuals, satisfying the axioms of Pollock's (1981) counterfactual logic **SS**.

In §4, I consider the differences between the exogenous intervention model for counterfactuals and the two leading alternative causal models, those of Hiddleston and Pearl. I argue that both models have limitations: Hiddleston struggles with some forward counterfactuals and Pearl struggles with backtracking counterfactuals. I also defend the choice of modeling with exogenous interventions and show how the main ideas of Pearl's and Hiddleston's models can both be incorporated into the exogenous intervention model by making extra assumptions on the model set-up and the selection function. In §5, I consider some issues which arise from relativizing truth conditions of counterfactuals to causal models, including backtracking interpretations of forward counterfactuals, counterlegal counterfactuals, and counterfactuals with counterfactual antecedents.

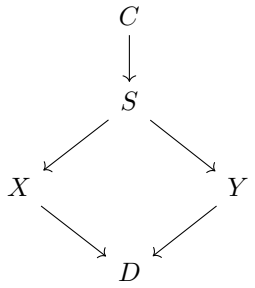
This paper therefore provides a causal model for counterfactuals which, unlike current causal frameworks, extends to logically complex antecedents and backtracking counterfactuals while providing a familiar counterfactual logic. It also serves to clarify the relationship between the causal and similarity models of counterfactuals as well as the relationship between different models of causal counterfactuals such as the interventionist approach of Pearl and the backtracking approach of Hiddleston. By showing that causal models are flexible enough to incorporate a wide range of complex counterfactuals and relate nicely to standard counterfactual logics, this paper aims to re-evaluate these two perceived weaknesses of causal modeling.

1 Causal Models

Consider a familiar example from the causal modeling literature, discussed in Pearl (2009): the firing squad. Here, a court is deciding whether to order the execution of a prisoner. If the court orders execution, then the captain sends a signal to two shooters, Shooter X and Shooter Y, who bring about the death of the prisoner. We can formalize this scenario as a causal model: we have five binary variables which take values 0 if the event does not occur and 1 if the event does occur and four structural equations describing the dependencies involved. We can write the components of the causal model as:

> **Variables** : the court orders execution (C), the captain sends a signal (S), Shooter X shoots (X), Shooter Y shoots (Y), prisoner dies (D)
>
> **Structural Equations** : $S = C$; $X = S$; $Y = S$; $D = X \vee Y$

We can also illustrate the causal dependencies in a graph:



The structural equations representing dependency relations allow us to use causal models to evaluate counterfactual sentences. We evaluate a counterfactual $A > C$ in a causal model by intervening in the model to set A true and seeing if this guarantees that C is true. Consider the counterfactual 'If X were to shoot, then the prisoner would die.' If we make an intervention on the causal model to set $X = 1$, then since $D = X \vee Y$, $D = 1$, so the prisoner must die; this renders the counterfactual true in this model.

To give a formal account of counterfactual truth conditions, we must define causal models more formally. A causal model $\mathcal{M} = (U, V, f_i)$ consists of a finite set of exogenous variables U, a finite set of endogenous variables $V = (V_1, ..., V_n)$, and a set of structural equations $F = (f_1, ..., f_n)$, where for each i, $v_i = f_i(pa_i, u_i)$, where $pa_i \in PA_i \subseteq V \backslash V_i$ is an assignment pa_i to the parents PA_i of V_i and $U_i \subseteq U$ is the unique minimal set of exogenous variables needed for f_i. Thus, each f_i tells us the value of the endogenous variable V_i given the value of V_i's parents PA_i and the exogenous variables U_i. The assignment of parents PA_i for V_i determines a graph $\mathcal{G}$ on V, which we assume is a directed acyclic graph (DAG). Since all endogenous variables have structural equations which depend on the variable's parents and exogenous variables, once we make an exogenous variable assignment $u \in U$, we fix the value of all endogenous variables, so the set of structural equations F forms a function from exogenous variables to endogenous variables, $F : U \to V$. Therefore, the values of the endogenous variables in a causal model are completely determined by the structural equations and the values of the exogenous variables.[3]

In the firing squad example, the only exogenous variable is the court ordering the execution (C); once the value of this variable has been settled, the values of all other variables are settled as well.[4] While the values of exogenous variables determine all other variables in a causal model, the significance of the distinction between exogenous and endogenous variables has often been ignored in causal models for counterfactuals.[5] In Pearl's model, for example, the distinction does

[3]For more details on the formal background to causal modeling, see Pearl (2009).

[4]Technically, C is an endogenous variable with no parents. However, we often think of these variables as being determined exogenously, so there is an exogenous variable U_C such that $C = U_C$.

[5]While Pearl uses exogenous variables in his original framework, these are left out in the more recent models of Hiddleston (2005), Kaufmann (2013), Santorio (2014), and Ciardelli et al. (2018).

not play any role: it does not matter whether the intervention we make is on an exogenous or an endogenous variable. Consider again the counterfactual 'If X were to shoot, then the prisoner would die.' In Pearl's interventionist approach to counterfactuals, hypothetically considering the antecedent replaces the structural equation $X = C$ with the structural equation $X = 1$; regardless of the fact that X is an endogenous variable, we can set it to $X = 1$ by breaking the causal laws of the model.

The distinction between exogenous and endogenous variables becomes significant when we try to incorporate backtracking counterfactuals, where the consequent is the cause of the antecedent. Consider the counterfactual 'If X were to shoot, then the captain signaled for it.' Intuitively, this counterfactual is true since, if the causal model is correct, X only shoots if the captain signaled to, so $X = 1$ only if $S = 1$. However, in a framework like Pearl's where we can intervene on any variable, this counterfactual need not be true. This is because, when we intervene directly on X to replace $X = S$ by $X = 1$, this does not change anything upstream from X, so the intervention does not guarantee that $S = 1$. In general, models where we can intervene directly on any variable cannot explain judgments about backtracking counterfactuals.[6]

In backtracking reasoning, we keep the laws, or structural equations, of the causal model the same, instead considering the changes on the exogenous variables which make the antecedent true. Preserving the interventionist intuition, we can consider changes to exogenous variables to be interventions.[7] In our example, C is the only exogenous variable, so the only way we can change any variables in the model while keeping the laws the same is by changing C. If we consider the exogenous interventions which set $X = 1$, our model tells us that X's decision to shoot is based solely on the signal S, and S, in turn, is based solely on C, so the only way to intervene within the model to set $X = 1$ is to set $C = 1$. This allows us to recover the desired truth conditions for the backtracking counterfactual 'If X were to shoot, then the captain signaled for it': intervening to set $X = 1$ involves setting $C = 1$, which sets $S = 1$, so the counterfactual is always true.

Note that, on this approach, the inclusion of exogenous variables is significant for counterfactual truth conditions: adding an extra exogenous variable, for example, can change the truth conditions of the backtracking counterfactuals. Suppose we think it is more accurate to attribute to X the possibility of shooting without receiving the signal. In this case, we should add an exogenous variable U_X to the causal model such that $X = S \vee U_X$ to account for this possibility, even if we consider the activation of U_X extremely unlikely. Exogenous variables

[6] Rips (2010) and Gerstenberg et al. (2013) provide experimental evidence supporting backtracking in counterfactual reasoning.

[7] This differs from the standard conception of an intervention following Pearl. Fisher (2017*a*), motivated by Pearl's concept of an intervention, argues that an intervention on an antecedent A requires making the variables in A independent of their parents. Here, an intervention on a variable in A can occur at a parent of that variable, as happens when we set $C = 1$ to fix $X = 1$, leaving the variables in A dependent on their parents. Thus, on Fisher's interpretation, an exogenous intervention does not properly count as an intervention in the model. I discuss these two notions of intervention in greater detail in §4.2.

like U_X are sometimes referred to as error terms because they introduce the possibility of outcomes deviating from the expected course of events. In this new causal representation of the situation, setting $X = 1$ can arise from setting either $U_X = 1$ or $S = 1$; the first intervention $U_X = 1$ does not guarantee that the captain gave the signal ($S = 1$) or that the court ordered the execution ($C = 1$). This shows how changing the exogenous variables included in a model can change judgments about counterfactuals: when U_X is not included, $X = 1 > S = 1$ is true, but when U_X is added to the model, $X = 1 > S = 1$ need not be true.

This discussion motivates the approach to counterfactuals I will define in the next section: $A > C$ is true if any intervention (or way of setting the exogenous variables in the model) which fixes A leads to C.

2 Exogenous Intervention Model

To draw the connection as closely as possible between causal models and the similarity-based theories of counterfactuals, I frame the discussion of causal models in terms of causal worlds. Pearl (2009) defines the notion of a causal world, but makes little use of the notion in his analysis, and the notion is largely left out of later causal models for counterfactuals. A causal world $(\mathcal{M}, u)$ is a causal model $\mathcal{M}$ paired with an assignment to all exogenous variables, $u \in U$. Since all endogenous variables are determined by an assignment $u \in U$, elements of U play the role of truthmakers for propositions of variable assignments, and we can associate propositions built from variable assignments with sets of worlds.

If $V_i = v_i$ is an endogenous variable assignment, this determines a set of possible worlds by $[V_i = v_i] = \{u \in U : F(u)_i = v_i\} \subseteq U$, so $u \in [V_i = v_i]$ iff $V_i = v_i$ is true when we plug u into the structural equations in $\mathcal{M}$. Since all variable assignments yield sets of possible worlds, any logical combination of variable assignments also determines a set of possible worlds as usual, where negation, conjunction and disjunction correspond to set-theoretic complementation, intersection, and union, respectively. As usual in possible world semantics, we refer to subsets of U as propositions, and the set of subsets of U, $\mathcal{P}(U)$, as the set of propositions. The truth conditions defined for counterfactuals will apply to all propositions, or sets of exogenous variable assignments; this definition is what allows us to extend the analysis of counterfactuals to antecedents with arbitrary logical complexity.

To see how this notion of causal worlds works, consider a modified version of the firing squad example where both X and Y are able to shoot without receiving the signal. Here, the causal graph is as above, but there are three exogenous variables, U_C, U_X, and U_Y, with structural equations $C = U_C, S = C, X = S \vee U_X, Y = S \vee U_Y$, and $D = X \vee Y$. In this case, there are eight possible worlds corresponding to the six possible assignments to the three exogenous variables. To see how complex propositions reduce to sets of worlds, consider the proposition 'The prisoner dies and either shooter X or shooter Y does not shoot.' We can see that there are only two worlds where this propositions is

true: $U_C = 0, U_X = 1, U_Y = 0$ and $U_C = 0, U_X = 0, U_Y = 1$.

To define the truth conditions associated with a counterfactual $A > C$, where A and C are propositions, we need to associate a world u and the antecedent A with a set of possible worlds over which we evaluate the consequent C; in similarity-based approaches, this is the set of closest A-worlds to u, determined by a selection function $f(A, u)$. The intuition behind causal counterfactual models is that the relevant set of A-worlds close to u is the set of worlds where we intervene in the causal model to make A true. This can be done by changing the structural equations, as in Pearl, or by changing the values of variables, as here and in Hiddleston. Here, I propose a characterization of the set of worlds formed by making an A-intervention on u; in §4, I argue that this set of interventions generalizes both Pearl's and Hiddleston's proposals. Furthermore, while the selection function used here is 'strict,' incorporating all intervened worlds, it is possible to define counterfactual semantics with a restricted selection function by imposing stronger similarity constraints within this model.

We consider an intervention i in u to be an intervention forcing A if it makes changes to the exogenous variables which are minimally necessary to ensure A. The exogenous variable assignments which force A are all the worlds where A is true, or the elements of $[A]$. However, we do not want to consider all assignments in which A is true; if C is independent of A, for example, intervening to fix A should not change C. For example, intervening to change the color of someone's shirt should not change their height. We want interventions to change only those variables which are necessary to produce A. These interventions, rather than being complete variable assignments (elements of U), will be partial variable assignments.

Suppose there are m exogenous variables, so $U = (U_1, ..., U_m)$, and let $S \subseteq \{1, ..., m\}$ be a subset of variables with complement S', so $U = U_S \times U_{S'}$ and for any $u \in U, u = u|_S \times u|_{S'}$. A restricted variable assignment r on S is an assignment $r \in U_S$; that is, a variable assignment for the restricted set of variables U_S. For a restricted variable assignment $r \in U_S$ and a world u, we define the world where we intervene on u by r as $u|r = r \times u|_{S'}$. This is the world where we change the values of u on S to the values r, but leave all other variables unchanged.

We then define the set of restricted variable assignments which force A in a world u:

$$R_u(A) = \{r : \exists S, r \in U_S \ \& \ u|r \in [A]\}.$$

This is the set of partial variable assignments such that imposing these variable assignments on the world u gives a world $u|r$ where A is true. As long as a proposition A is possible, or has some world $w \in [A]$ making it true, $R_u(A) \neq \emptyset$ since $w \in R_u(A)$ with $S = \{1, ..., m\}$; every element $w \in [A]$ is in $R_u(A)$ for any u. However, as motivated above, we don't want all elements of $[A]$ to be interventions on A, so we must restrict the set $R_u(A)$.

We want to restrict $R_u(A)$ to just include the variable changes which are necessary to bring about A. If r is such a variable change, then fixing any other variables in addition to those fixed by r would not change the value of A; this

means any extension of r to variables not fixed by r is also an element of $R_u(A)$. This motivates defining an order $\leq$ on $R_u(A)$. Suppose $r_1, r_2 \in R_u(A)$ assign variables S_1 and S_2. We say $r_1 \leq r_2$ iff r_2 is an extension of r_1, so $S_1 \subseteq S_2$ and $r_2|_{S_1} = r_1$.

We can now define the set of interventions which force A, $I_u(A)$, as the $\leq$-minimal elements of $R_u(A)$:

$$I_u(A) = \{i \in R_u(A) : \nexists r \in R_u(A), r \neq i, r \leq i\}.$$

We then define the truth conditions for a counterfactual: a counterfactual $A > C$ is true in a world u if C is true when we make all interventions from $I_u(A)$ on u. Thus, the set of worlds where a counterfactual $A > C$ is true is as follows:

$$[A > C] = \{u \in U : \forall i \in I_u(A), u|i \in [C]\}.$$

To see that these truth conditions provide intuitive results, recall the modified firing squad example from above with exogenous variables U_C, U_X, and U_Y and structural equations $C = U_C, S = C, X = S \vee U_X, Y = S \vee U_Y$, and $D = X \vee Y$. Suppose that, in the actual world, the court does not order execution and neither X nor Y choose to shoot, so $(U_C, U_X, U_Y) = (0, 0, 0)$. Consider the counterfactual 'If X were to shoot, the prisoner would die.' Here, the relevant interventions are $U_X = 1$ and $U_C = 1$; in both cases, $X = 1$, so $D = 1$, so the counterfactual is true. Now consider the backtracking counterfactual 'If X or Y were to shoot, the captain must have signaled.' The relevant interventions are $U_X = 1$, $U_Y = 1$, and $U_C = 1$, and under the interventions $U_X = 1$ and $U_Y = 1$, $C = 0$, so the counterfactual is false. This is intuitively correct: since it is possible that X or Y decides to shoot without receiving the signal, X or Y shooting does not entail that the captain signaled. These examples show how the truth conditions defined here offer reasonable predictions for both forward and backtracking counterfactuals.

3 Logic of Exogenous Intervention Models

Similarity-based models for counterfactuals rely on selection functions $f(A, u) : \mathcal{P}(U) \times U \to \mathcal{P}(U)$, which assign a world u and antecedent A to a set of closest relevant A-worlds to u. The exogenous intervention model defines a selection function by $f(A, u) = \{u|i : i \in I_u(A)\}$. The logic for similarity-based models of counterfactuals built from selection functions is well-understood; restrictions on the selection function f correspond to axioms for the conditional $>$.[8] The best-known logic for counterfactuals is Lewis's **VC**, which corresponds to six axioms on selection functions:

CS1: if $w \in f(A, u)$, then $w \in [A]$
CS2: if $u \in [A]$, then $f(A, u) = \{u\}$

[8] See the classic text of Lewis (2013) or the recent surveys of Nute and Cross (2001) and Arlo-Costa (2019).

CS3: if $f(A,u) = \emptyset$, then $f(B,u) \cap [A] = \emptyset$
CS4: if $f(A,u) \subseteq [B]$ and $f(B,u) \subseteq [A]$, then $f(A,u) = f(B,u)$
CS5: if $f(A,u) \cap [B] \neq \emptyset$, then $f(A \wedge B,u) \subseteq f(A,u)$
CS6: $i \in [A > C]$ iff $f(A,u) \subseteq [C]$

However, many authors have recommended weaker logics than **VC**. Pollock (1981), for example, recommends a logic **SS**, where we replace **CS5** by **CS5′**:

CS5′: $f(A \vee B,u) \subseteq f(A,u) \cup f(B,u)$

The selection function for exogenous intervention models defined above satisfies the axioms for Pollock's logic **SS**. We verify the satisfaction of these six axioms below:

CS1: if $w \in f(A,u)$, then $w \in [A]$

Proof. Suppose $w \in f(A,u)$, so $w = u|i$ for some $i \in I_u(A)$. Since $i \in R_u(A)$, $u|i \in [A]$ by the definition of $R_u(A)$, so $w \in [A]$. □

CS2: if $u \in [A]$, then $f(A,u) = \{u\}$

Proof. If $u \in [A]$, then the empty intervention i, which changes no exogenous variables, is in $R_u(A)$ since $u|i = u \in R_u(A)$. Since $i \leq r$ for every other possible intervention $r \in R_u(A)$, i is the unique minimal element in $R_u(A)$ and the only element in $I_u(A)$. Since $f(A,u) = \{u|i : i \in I_u(A)\}, f(A,u) = \{u|i\} = \{u\}$. □

CS3: if $f(A,u) = \emptyset$, then $f(B,u) \cap [A] = \emptyset$

Proof. If $f(A,u) = \emptyset$, then $I_u(A) = \emptyset$, so $R_u(A) = \emptyset$. Since $[A] \subseteq R_u(A)$, $[A] = \emptyset$, so $f(B,u) \cap [A] = \emptyset$. □

CS4: if $f(A,u) \subseteq [B]$ and $f(B,u) \subseteq [A]$, then $f(A,u) = f(B,u)$

Proof. Suppose $f(A,u) \subseteq [B]$ and $f(B,u) \subseteq [A]$. To show that $f(A,u) \subseteq f(B,u)$, we must show that, for all $i \in I_u(A)$, there is some $j \in I_u(B)$ such that $u|i = u|j$. Suppose $i \in I_u(A)$. Since $f(A,u) \subseteq [B]$, $u|i \in [B]$, so $i \in R_u(B)$. Then there is a $j \in I_u(B)$ such that i extends j. But since $j \in I_u(B)$ and $f(B,u) \subseteq [A]$, $u|j \in [A]$, so $j \in R_u(A)$. This means there is an $i' \in I_u(A)$ such that j extends i'. But since i and i' are both $\leq$-minimal elements and $i' \leq j \leq i$, $i = i' = j$, so $u|i = u|j$. Since we have shown $\forall i \in I_u(A), \exists j \in I_u(B)$ such that $u|i = u|j$, we have shown that $f(A,u) \subseteq f(B,u)$. The proof that $f(B,u) \subseteq f(A,u)$ is parallel, showing that $f(A,u) = f(B,u)$. □

CS5′: $f(A \vee B,u) \subseteq f(A,u) \cup f(B,u)$

Proof. Suppose $u|i \in f(A \vee B, u)$, where $i \in I_u(A \vee B)$. Since $u|i \in [A \vee B]$ by **CS1**, $u|i \in [A]$ or $u|i \in [B]$. Suppose $u|i \in [A]$. Then $i \in R_u(A)$, so there is some $j \in I_u(A)$ such that i extends j. Since $j \in I_u(A)$, $u|j \in [A] \subseteq [A \vee B]$, so $j \in R_u(A \vee B)$. This means there is some $i' \in I_u(A \vee B)$ such that j extends i'. But since $i' \leq j \leq i$ and i and i' are both $\leq$-minimal, $i = i' = j$, so $\exists j \in I_u(A)$ such that $u|i = u|j$, so $u|i \in f(A, u) \cup f(B, u)$. If $u|i \in [B]$, a parallel proof shows that $u|i \in f(B, u) \subseteq f(A, u) \cup f(B, u)$. Therefore, $f(A \vee B, u) \subseteq f(A, u) \cup f(B, u)$. □

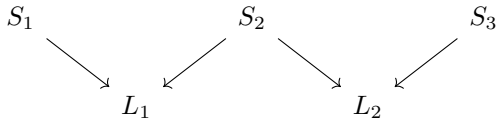
CS6: $i \in [A > C]$ iff $f(A, u) \subseteq [C]$

Proof. Follows immediately from the definition of $[A > C]$ in §2. □

Note that the exogenous intervention model does not satisfy Lewis's logic **VC** as it admits counterexamples to **CS5** and the corresponding logical principle:

$$(A > C) \wedge \neg(A > \neg B) \Rightarrow (A \wedge B) > C.$$

The counterexample to this is the same as found in Pollock and translated to causal models in Hiddleston. Suppose three switches S_1, S_2, and S_3 control two lights L_1 and L_2 with structural equations $L_1 = S_1 \vee S_2$ and $L_2 = S_2 \vee S_3$. The causal diagram for this model is as follows:



Suppose all three switches are off ($S_i = 0$) and, consequently, both lights are off ($L_i = 0$). The counterfactual 'If L_2 were on, S_1 would be off' is true since both interventions which set $L_2 = 1$, $S_2 = 1$ and $S_3 = 1$, leave S_1 fixed at 0. Additionally, it is not the case that 'If L_2 were on, L_1 would be on' since setting $S_3 = 1$ is an intervention which fixes $L_2 = 1$ without setting $L_1 = 1$. However, it is not the case that 'If L_1 and L_2 were on, S_1 would be off' since $(S_1, S_3) = (1, 1)$ is a minimal intervention setting the antecedent true. This provides a counterexample to the logical principle corresponding to **CS5**, showing that the exogenous intervention model does not validate Lewis's semantics **VC** without additional restrictions on the selection function.

This logic differs from alternative logics proposed for causal models. Briggs, for example, offers a logic for an extension of Pearl's theory which does not satisfy modus ponens. **SS**, however, satisfies modus ponens, which is a consequence of strong centering (**CS2**). Pearl and Halpern also show that Pearl's framework corresponds to the logic of Lewis's **VC** for antecedents restricted to conjunctions of variable assignments, while the counterexample above shows that **VC** is stronger than the logic of the fully general exogenous intervention model. As I discuss in §4.2, natural extensions of Pearl's framework to include disjunctions of variable assignments also invalidate **CS5**, failing to correspond to Lewis's **VC**. The characterization of the logic of causal models in terms of selection

functions also differs from other approaches, yielding an easier comparison with similarity-based models and more familiar proofs.

4 Comparison to Other Causal Models

The exogenous intervention model is designed to generalize both Hiddleston's and Pearl's causal models of counterfactuals. It combines elements of both models; for example, it makes use of the exogenous variables fundamental to Pearl's account while focusing on changes to variables rather than changes to structural equations, following Hiddleston. In this section, I show how Hiddleston's (§4.1) and Pearl's (§4.2) frameworks fit in the exogenous intervention model, defending the modeling choices made in this paper.

Both Hiddleston's and Pearl's models motivate certain restrictions on the semantics defined in §2. These restrictions arise from imposing an order relation $\leq$ on $I_u(A)$ and requiring for $A > C$ to be true that C is true under all $\leq$-minimal interventions in $I_u(A)$, rather than all interventions in $I_u(A)$. When considering restrictions to the semantics from §2, we can think of $I_u(A)$ as the set of all causally relevant A worlds close to u; the further restriction, $\leq$, can come from any other ordering on variables deemed relevant for the counterfactual semantics. We could consider restrictions where the selection function consists of a single closest world, yielding Stalnaker's semantics **C2**, or restrictions which validate **CS5** to get Lewis's logic **VC**, including restrictions which violate the Limit Assumption and require a System of Spheres model. Here, we only consider two restrictions coming from two of the major causal theories of counterfactuals, both of which yield the same logic **SS**.[9]

4.1 Hiddleston's Theory

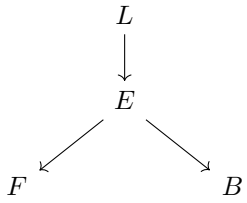
Hiddleston evaluates a counterfactual $A > C$ at u by considering whether C is true in all models which are 'minimal breaks' from the model in u; these models leave the structural equations unchanged and change variables upstream from A in a way which involves making the smallest necessary changes to variables while leaving the most variables independent of A intact. We can think of the set of possible breaks as $R_u(A)$ and Hiddleston's minimal change requirement as an ordering $\leq_H$ on $R_u(A)$. After discussing some foundational differences, I will show how we can define an ordering $\leq_H$ such that the set of $\leq_H$-minimal elements of $R_u(A)$ forms a subset of $I_u(A)$.

Hiddleston's theory follows the set-up of §1 with two fundamental differences: he considers all variables as endogenous and he allows for indeterministic structural equations such as $\Pr(Y = y|X = x) = p$. While he does not include an explicit set of exogenous variables, it is implicit that any variable which has

[9]The proofs that the restrictions by $\leq_H$ and $\leq_P$ satisfy the axioms for **SS** follow the proofs given in §3. The counterexample to **CS5** from §3 applies to Hiddleston's theory, as discussed in his paper, and a counterexample to **CS5** for the extension of Pearl's theory to disjunctive antecedents is given in §4.2.

no parents in the graph $\mathcal{G}$ over V is exogenous; it can be freely set with no constraints from its structural equation because structural equations only involve the parents of a variable. Thus, for any variable V_i such that $PV_i = \emptyset$, we can add an exogenous variable U_i such that the structural equation for V_i is $V_i = U_i$.

The issue of indeterministic structural equations is a little more difficult to resolve, but Pearl provides adequate machinery to handle indeterminacies. To see how this works, consider Hiddleston's ceremonial cannon example. Here, one lights a fuse (L), which has a 95% chance of setting off an explosion (E), which causes a flash (F) and a bang (B). The structural equations are $\Pr(E = 1|L = 1) = 0.95$, $\Pr(E = 1|L = 0) = 0$, $F = E$, and $B = E$ with causal graph:



On Pearl's theory, indeterminacies are handled by exogenous error variables rather than indeterministic structural equations. In this example, we would add an error variable U'_E representing 'unspecified inhibiting abnormalities' and replace Hiddleston's structural equation for E by $E = L \wedge \neg U'_E$, meaning that E is activated when L is activated and L is not inhibited by U'_E.[10] The fact that lighting the fuse leads to an explosion 95% of the time corresponds to the fact that there is a 5% chance the error variable U'_E is activated, or $\Pr(U'_E) = .05$. Removing indeterminacies from Hiddleston's structural equations is important because Hiddleston uses indeterministic structural equations to justify intervening on endogenous variables. For example, if we want to intervene to set $E = 0$, Hiddleston argues that we can do this by changing the endogenous variable E directly without changing its parent L; on an exogenous intervention theory, however, we can only change E by changing an exogenous variable, either L or U'_E.

This discussion shows how we can replace Hiddleston's indeterministic model with only endogenous variables by a model with deterministic structural equations and exogenous variables. This can be accomplished by adding exogenous variables corresponding to the endogenous variables with no parents and replacing indeterministic structural equations by deterministic equations with an exogenous error term. In the model above, we keep the same endogenous variables and the same causal graph; we just add two exogenous variables U_L corresponding to the only vertex without a parent and U'_E corresponding to the error associated with the statistical law. The new structural equations are then $L = U_L$, $E = L \wedge \neg U'_E$, $F = E$, and $B = E$.

Given the different representations for causal models, the predictions of Hiddleston differ from those of this paper. The use of indeterministic structural

[10] For Pearl's discussion of error variables in Boolean models, see Pearl (2009, p. 29).

equations in Hiddleston, for example, makes it difficult to interpret forward counterfactuals. Consider a world where the cannon is not lit and where we want to evaluate the counterfactual 'If the cannon were lit, then an explosion would happen.' In the exogenous intervention model, this is true in 95% of worlds (where $U'_E = 0$) and false in the other 5%, but in Hiddleston's model, this is predicted false in general because there is always a possible outcome where the cannon is lit but an explosion does not occur.

However, Hiddleston does offer a restriction on the semantics for counterfactuals which may be useful for us. While the exogenous intervention model in §2 evaluates counterfactuals over all relevant minimal changes to the exogenous variables, Hiddleston further restricts to those 'causal breaks' which are 'as minor and as late as is lawfully possible' (Hiddleston, 2005, p. 643). While the condition of minimality is enforced in §2 to rule out unnecessary interventions, we considered all interventions rather than those that are as late as possible in the causal process. To enforce this additional requirement, we can demand that the set of variables independent of the antecedent remains as intact as possible, so that we consider only those interventions which minimally change the variables independent of the antecedent. Let $V(u|i)$ be the endogenous variable assignment produced by intervention i in world u, Z the set of variables which are not descendants of any variables in A, and $V(u|i) \cap Z$ the set of variable assignments in $V(u|i)$ which are in Z and equal to those in the world u. We can define an order on $I_u(A)$, $\leq_H$, by saying $i \leq_H i'$ iff $V(u|i') \cap Z \subseteq V(u|i) \cap Z$.

We can then define the counterfactual $A > C$ as true at u iff C is true under all $\leq_H$-minimal interventions from $I_u(A)$. This offers a restriction on the semantics in §2 which yields different truth conditions for counterfactuals. To see that the new truth conditions are different, consider the above example where the ceremonial cannon was lit, exploded, and the flash and bang occurred in the actual world. Consider the counterfactual 'If the flash hadn't occurred, the cannon was still lit.' On the strict semantics from §2, this counterfactual is false. There are two minimal interventions which could turn off the flash, one where the cannon isn't lit ($U_L = 0$) and the other where an error prevents the lit cannon from exploding ($U'_E = 1$). Since the cannon is lit in one of these but not the other, the counterfactual is false. On Hiddleston's theory and the restricted semantics here, however, the intervention $U'_E = 1$ leaves more independent variables intact (namely L), so it is the only relevant intervention, meaning the counterfactual is true. Thus, if we modify Hiddleston's theory to fit into the exogenous intervention semantics, we can recover a restricted counterfactual semantics with slightly different truth conditions than found in §2.

4.2 Pearl's Theory

Pearl argues that a counterfactual $A > C$ is true if an intervention to produce A entails C, where we intervene on A by replacing the structural equation for A with $A = 1$ rather than changing the values of exogenous variables. Pearl's model is limited insofar as it cannot handle logically complex antecedents or

backtracking counterfactuals. However, Pearl draws on extensive evidence from the theory of causal inference to justify these interventions on structural equations as the correct representation of counterfactual intervention. Pearl intentionally avoids backtracking in counterfactual reasoning because backtracking can lead one to ignore confounders and mistake correlation for causation.

Consider the case of monetary policy, where a central banker considers lowering interest rates to increase output and inflate prices. Typically, the monetary policy decision is made based on economic fundamentals, making the decision endogenous. Suppose a central banker ignores the economic fundamentals and reasons: if I were to lower interest rates, then economic fundamentals would be as they usually are when the central bank lowers interest rates, and output and prices would therefore increase. This backtracking reasoning is clearly erroneous and confuses the correlation of monetary policy decisions and economic effects with a causal effect of monetary policy on the economy. Instead, Pearl argues, we should evaluate the consequences of a monetary policy decision by taking the fundamentals as given, intervening to set the interest rates to a certain level, and seeing how (if at all) this affects the economy. Pearl's approach to interventions resolves the backtracking problem: the monetary policy decision can remain endogenous and we can (correctly) consider an intervention as something which does not change the background fundamentals.

This is a serious obstacle to implementing a theory of counterfactuals which can handle backtracking counterfactuals: in many decision environments, backtracking seems inappropriate. However, we can resolve this in the exogenous intervention model by adding exogenous variables to our model. In the monetary policy example, we can treat an intervention not as a break in the structural equations, but rather as an exogenous variable which influences the interest rate directly without influencing the fundamentals. We can justify adding this exogenous variable because, in order for there to be a real possibility of intervening on an endogenous variable, there must be some way to change the variable regardless of the value of its parents. This is precisely what an intervention is, and also precisely what an exogenous variable represents. One way of thinking of the additional exogenous variable is as an error term representing all possible ways of influencing the endogenous variable not covered by the parent variables. Since causal models almost never list all possible influences, we expect such an error variable to exist, even if we consider it negligible in most modeling circumstances.

When considering monetary policy, for example, any input to the interest rate decision which does not come from economic fundamentals can be considered part of the exogenous error term. While in most circumstances we consider this exogenous input to the interest rate decision negligible, we can certainly add it to our model. Economists, for example, have tried to isolate situations in which this exogenous variable is activated by identifying cases when central banks make decisions which deviate from what is expected based on the economic fundamentals.[11] Models which consider such exogenous interventions a

[11] One way of measuring this in the US is by noting when the Fed funds rate deviates from

salient possibility, such as models where the economy can be subject to a 'monetary policy shock,' even explicitly include an exogenous variable influencing interest rate decisions.[12] Therefore, while Pearl would consider an intervention on interest rates a change to the structural equations, the exogenous intervention model interprets the possibility of such an intervention as an exogenous variable influencing interest rates. The fact that economists estimate this exogenous effect on interest rates and incorporate exogenous variables representing it in their models serves as evidence for interpreting such an intervention as an exogenous variable rather than a change to the structural equations.

We can also get the same predictions as Pearl's semantics in an exogenous intervention model if we add an exogenous error term to every endogenous variable in the model and define an ordering $\leq_P$ on $I_u(A)$ which makes sure we only change the exogenous variables governing the variables in A. To add the exogenous variables, we assume each endogenous variable V_i has an exogenous variable $U_i = V_i \cup \{\text{OFF}\}$, where V_i is determined according to its original structural equation when $U_i = \text{OFF}$ and $V_i = U_i$ otherwise. Activating the exogenous variable can be very unlikely, i.e., $\Pr(U_i = \text{OFF}) \approx 1$; what matters is that the possibility is included in the model.

To define an ordering which ensures that we only change the exogenous variables governing variables in A, we will define a notion of distance between an intervention and an antecedent which is minimized by exogenous variables U_i which directly affect variables V_i. For an exogenous variable U_i which enters the graph $\mathcal{G}$ at vertex V_i, for any vertex V_j, let $d(U_i, V_j)$ be the graph theoretical distance between V_i and V_j. For an antecedent A which references variables V_A and for an intervention i on variables U_S, define

$$d(i, A) = \sum_{s \in S} \min_{V_j \in V_A} d(U_s, V_j).$$

For two interventions $i, i' \in I_u(A)$, we say $i \leq_P i'$ iff $d(i, A) \leq d(i', A)$ and a counterfactual $A > C$ is true at u iff C is true under all $\leq_P$-minimal interventions from $I_u(A)$.

We can now see that this gives us Pearl's semantics for counterfactuals. Suppose that the antecedent is a variable assignment $V_i = v_i$. The only non-empty $\leq_P$-minimal interventions are those which change the exogenous variables at V_i since we know $U_i = v_i$ is such an assignment; any such intervention leaves the variables upstream from V_i constant and coincides with Pearl's intervention to set the structural equation for V_i to $V_i = v_i$.[13] Similarly, for conjunctions of variable assignments, the intervention which changes the exogenous variables for each conjunct is $\leq_P$-minimal, and any other minimal intervention coincides with this on endogenous variables, again coinciding with the predictions of Pearl.

futures on the Fed funds rate. See Kuttner (2001).

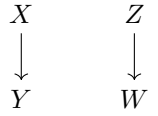
[12] See, for example, Christiano et al. (2005).

[13] Note that when $V_i = v_i$ is true in a world u, $I_u(A)$ consists of only the empty intervention, so the $\leq_P$-minimal element of $I_u(A)$ is the empty intervention. This may leave exogenous variables at V_i unactivated, but activating them wouldn't change any other variables in the model. The predictions in this case, therefore, also coincide with those of Pearl.

Thus, for all the antecedents which Pearl considers, his semantics arises as a restricted version of the exogenous intervention theory.

This version of Pearl's theory extends the theory to logically complex antecedents, for example, to disjunctions of variable assignments. If we have an antecedent $(V_i = v_i) \vee (V_j = v_j)$, the interventions $U_i = v_i$ and $U_j = v_j$ would each be minimal, corresponding to two possible ways of replacing the structural equations: replacing the structural equation for V_i with $V_i = v_i$ and replacing the structural equation for V_j with $V_j = v_j$. Note that this differs from Santorio's (2014) extension of Pearl's framework to disjunctions, since he also considers the possibility of replacing both structural equations simultaneously. In the framework here, the conjoined variable assignment $(U_i, U_j) = (u_i, u_j)$ is an extension of both individual assignments, so is not an element of $I_u(A)$. Considering this conjunctive variable assignment for disjunctions would violate **CS5′**, and Ciardelli et al. (2018) provide experimental evidence that people do not consider the combined intervention when evaluating counterfactuals with disjunctive antecedents.[14]

While Pearl claims that his logic satisfies the axioms of Lewis's logic **VC** for antecedents limited to conjunctions of variable assignments, this claim has been controversial. Halpern (2013), for example, argues that Pearl's proof is incorrect because it ignores logical consequences of incorporating disjunctions into Lewis's system, though he offers a new proof resolving this issue. Once we incorporate disjunctions into Pearl's framework using exogenous interventions, we can see that Pearl's logic satisfies Pollock's **SS** rather than Lewis's **VC**. Consider the following graph, for example:



with structural equations $Y = X$ and $W = Z$. Let $T = (Y \vee W) \wedge \neg(Y \wedge W)$. Recall that **CS5** corresponds to the logical principle $(A > C) \wedge \neg(A > \neg B) \Rightarrow (A \wedge B) > C$. In this case, if we assume all variables have value 0 in the actual world, we can see that on Pearl's semantics, $(X \vee Z) > T$ is true, $\neg((X \vee Z) > \neg Y)$ is true, but $((X \vee Z) \wedge Y) > T$ is false. The final counterfactual is false because the intervention $(U_Z, U_Y) = (1, 1)$ ensuring $Z \wedge Y$ is a $\leq_P$-minimal intervention (both Z and Y are in the antecedent) which makes T false. Therefore, this extension of Pearl's semantics to disjunctions violates **CS5**, yielding **SS** rather than **VC**.

[14] Ciardelli et al. (2018) offer an interesting semantics for counterfactuals which considers conjunctions of possible interventions in the set of relevant interventions, eliminating them for disjunctive antecedents by a 'lifting' of the counterfactual semantics into inquisitive semantics. It is unclear whether this approach can be evaluated in terms of the conditions on selection functions in §3.

5 Model Selection and Limitations of Causal Modeling

One issue that arises for causal modeling of counterfactuals is the selection of the correct causal model for analysis. As the vast differences in predictions of Hiddleston's and Pearl's models indicate, a decision as small as whether to include an exogenous variable can have large effects on predictions for the truth conditions of counterfactuals. This issue arose in the original firing squad case, where whether we include an exogenous variable governing Shooter X's ability to shoot without a signal from the captain determines the truth values of counterfactuals like 'If X shoots, then the captain gave the signal.' Certainly, it is possible in some sense for X to shoot without command, but is this possibility relevant or negligible?

While the issue of model selection can leave the truth values of counterfactuals underdetermined, this is not always problematic. There are many cases, such as the case of monetary policy intervention considered in §4.2, where agents disagree about the causal structure and the variables which are and are not subject to exogenous shocks. The causal modeling framework therefore provides a clear explanation of how people can disagree about the truth value of a counterfactual: disagreement about the underlying causal structure. Disagreement about the correct causal model can also explain a major type of disagreement identified in the literature on counterfactuals: backtracking readings of forward counterfactuals.

Consider the situation from Jackson (1977), also discussed in Khoo (2017): your friend Smith is on top of a building about to jump, but steps off. You say, 'If Smith had jumped, he would have died,' which appears true. Your friend Beth, however, hears you say this and disagrees, arguing that Smith has no desire to die, so if he had jumped, there would have been a net or something else intervening to prevent his death, so she claims, 'If Smith had jumped, he would not have died.' The first prediction about the counterfactual is a forward reading, while Beth's interpretation is a backtracking reading.

The original speaker probably had a simple causal model in mind: jumping off a building leads one to die, so there is one exogenous variable U_J, $J = U_J$, and death is determined by J, $D = J$. However, Beth proposes a different causal model: there is the possibility that some condition, like a net, will prevent jumping from causing death, and this is likely the case in the actual world due to Smith's psychology. Now, we have two exogenous variables, U_J and U'_D, where $J = U_J$ and $D = J \wedge \neg U'_D$. Beth claims that $U'_D = 1$, so the counterfactual $J = 1 > D = 1$ is false on her model of the world, even though it was true on the original model. The other examples of backtracking counterfactuals in the literature can be handled similarly by describing different causal models for the forward and backtracking readings of the counterfactuals. This interpretation of backtracking counterfactuals as arising from disagreement about the causal model differs from the 'historical modality' account of backtracking found in Khoo (2017) and counters Lee's (2015) thesis that the ambiguity of counterfac-

tuals exhibited in forward and backtracking readings requires separate causal models for intervention and extrapolation.[15]

Another issue which arises for model selection is that some models will render an antecedent impossible, so there is no way to evaluate the truth value of the counterfactual. Consider again the firing squad case where we now seek to evaluate the counterfactual 'If someone replaced all the real bullets with rubber bullets, then the prisoner would have died.' All causal models considered for the firing squad provide no way to evaluate this counterfactual, since the antecedent does not correspond to variables considered in the model. Intervening to set the antecedent true would require breaking the causal law $D = X \vee Y$, which is impossible. Thus, all the causal models we considered treat the counterfactual as either not evaluable or vacuously true, despite the fact that it is intuitively false. If we had a model where this possibility was accounted for, perhaps by adding an error term U'_D which is activated when something (like rubber bullets) prevents X or Y shooting from causing the death of the prisoner, $D = (X \vee Y) \wedge \neg U'_D$, then we would have had no problem evaluating this counterfactual.

However, not all counterfactuals can be evaluated by choosing an appropriate causal model. Some counterfactual antecedents would never be true in any reasonable causal model, such as 'If turning off the sprinklers caused it to rain,...' or 'If Shooter X was subversive and did the opposite of what the captain ordered,...' These, as well as the case above with the rubber bullets, are often called counterlegal counterfactuals since they require breaking the laws of the causal model. While some counterlegal counterfactuals can be handled simply by adding new variables, others, such as those requiring us to change a structural equation directly, cannot be evaluated simply in a causal model.[16]

This problem of counterlegal counterfactuals also makes it difficult to evaluate counterfactual antecedents in causal models. Since we defined a set of possible worlds corresponding to counterfactuals, $[A > C]$, and the evaluation of a counterfactual only relied on the antecedent and the consequent being propositions, or sets of possible worlds, we might expect to obtain good results for counterfactual antecedents. However, this is not the case. Suppose one has a test (T) for a disease (D) which gives a positive result if the disease is present with a chance of false positives (U_T), so $T = D \vee U_T$. Consider the counterfactual 'If you were to have the disease when the test is positive, then you would have the disease,' represented $(T > D) > D$, in a world where you don't have the disease and the test is negative. Formally, $T > D$ is true when any intervention which fixes T fixes D, which is only true if D is true since $U_T = 1$ is always a possible intervention. Thus, setting D true is the only relevant intervention which fixes $T > D$, which ensures that D is true, so the counterfactual $(T > D) > D$ is

[15] Lee (2017, p. 90) offers another example, *Nuclear*, motivating the need for a dual theory of intervention and extrapolation in causal models. In this example, no variable changes make a given antecedent A true, but Pearl-style intervention can make the antecedent true. However, the modification of Pearl's theory in §4.2 where we associate interventions with exogenous variables resolves this division between intervention and extrapolation.

[16] Fisher (2017*b*) provides one possible solution to this problem, proposing a way to go from a given causal model $\mathcal{M}$ where the antecedent is impossible to a set of closest ('minimally illegal') causal models where the antecedent is possible.

formally true. However, intuitively, we expect the counterfactual to be false, since the fact that a test accurately indicates disease does not entail that someone has that disease. This is because the antecedent is not asking us to change variables within the model, but rather to change the causal model. The antecedent $T > D$ is true if we remove the error variable U_T from the model, or if the test has no false positives. However, the given causal model does not offer a way to incorporate such a counterlegal antecedent, making evaluation of this counterfactual difficult in the causal model.

This section shows that, in a causal modeling approach, the truth conditions of counterfactuals are closely tied to the causal model selected for analysis. Small changes to a causal model, such as adding or taking away error terms, can change the predictions for a large number of counterfactual sentences. This sensitivity has some benefits, allowing us to explain disagreements in counterfactual judgments and the possibility of forward and backtracking readings for the same counterfactuals. However, it also comes with some costs, making it challenging to offer predictions for counterfactuals with antecedents that require changing the causal model, such as counterlegal counterfactuals and counterfactuals with counterfactual antecedents.

6 Conclusion

In this paper, I introduced a causal model for counterfactuals, the exogenous intervention model, which incorporates logically complex antecedents and yields a familiar counterfactual logic, Pollock's **SS**. This model predicts that a counterfactual $A > C$ true in a causal world u if C is true in all worlds formed by intervening to set A true. This differs from other causal models which consider only restricted antecedents and which have less familiar logical properties. The exogenous intervention model also generalizes the causal models of both Pearl and Hiddleston, incorporating an analysis of both forward and backtracking counterfactuals. The relativization of truth conditions for counterfactuals to causal models can also explain disagreement about counterfactuals and how we get forward and backtracking interpretations of the same counterfactuals. However, the semantics remains limited for antecedents which require amending the causal model, such as counterlegal counterfactuals and counterfactuals with counterfactual antecedents.

References

Arlo-Costa, H. (2019), The logic of conditionals, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Metaphysics Research Lab, Stanford University.

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

Campos, N. F., Coricelli, F. and Moretti, L. (2019), 'Institutional integration and economic growth in Europe', *Journal of Monetary Economics* **103**, 88–104.

Christiano, L. J., Eichenbaum, M. and Evans, C. L. (2005), 'Nominal rigidities and the dynamic effects of a shock to monetary policy', *Journal of Political Economy* **113**(1), 1–45.

Ciardelli, I., Zhang, L. and Champollion, L. (2018), 'Two switches in the theory of counterfactuals', *Linguistics and Philosophy* **41**(6), 577–621.

Fisher, T. (2017*a*), 'Causal counterfactuals are not interventionist counterfactuals', *Synthese* **194**(12), 4935–4957.

Fisher, T. (2017*b*), 'Counterlegal dependence and causations arrows: causal models for backtrackers and counterlegals', *Synthese* **194**(12), 4983–5003.

Galles, D. and Pearl, J. (1998), 'An axiomatic characterization of causal counterfactuals', *Foundations of Science* **3**(1), 151–182.

Gerstenberg, T., Bechlivanidis, C. and Lagnado, D. A. (2013), Back on track: Backtracking in counterfactual reasoning, *in* 'Proceedings of the Annual Meeting of the Cognitive Science Society', Vol. 35, pp. 2386–2391.

Glymour, C. N. (2001), *The mind's arrows: Bayes nets and graphical causal models in psychology*, MIT press.

Gopnik, A. and Schulz, L. (2007), *Causal learning: Psychology, philosophy, and computation*, Oxford University Press.

Halpern, J. Y. (2000), 'Axiomatizing causal reasoning', *Journal of Artificial Intelligence Research* **12**, 317–337.

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

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

Jackson, F. (1977), 'A causal theory of counterfactuals', *Australasian Journal of Philosophy* **55**(1), 3–21.

Kaufmann, S. (2013), 'Causal premise semantics', *Cognitive Science* **37**(6), 1136–1170.

Khoo, J. (2017), 'Backtracking counterfactuals revisited', *Mind* **126**(503), 841–910.

Kuttner, K. N. (2001), 'Monetary policy surprises and interest rates: Evidence from the Fed funds futures market', *Journal of Monetary Economics* **47**(3), 523–544.

Lee, K. Y. (2015), Causal models and the ambiguity of counterfactuals, *in* 'International Workshop on Logic, Rationality and Interaction', Springer, pp. 220–229.

Lee, K. Y. (2017), 'Hiddlestons causal modeling semantics and the distinction between forward-tracking and backtracking counterfactuals', *Studies in Logic* **10**(1), 79–94.

Lee, R. S. (2013), 'Vertical integration and exclusivity in platform and two-sided markets', *American Economic Review* **103**(7), 2960–3000.

Lewis, D. (2013), *Counterfactuals*, John Wiley & Sons.

Nute, D. and Cross, C. B. (2001), Conditional logic, *in* 'Handbook of Philosophical Logic', Springer, pp. 1–98.

Pearl, J. (2009), *Causality*, Cambridge university press.

Pollock, J. L. (1981), 'A refined theory of counterfactuals', *Journal of Philosophical Logic* pp. 239–266.

Rips, L. J. (2010), 'Two causal theories of counterfactual conditionals', *Cognitive Science* **34**(2), 175–221.

Santorio, P. (2014), Filtering semantics for counterfactuals: Bridging causal models and premise semantics, *in* 'Semantics and Linguistic Theory', Vol. 24, pp. 494–513.

Sloman, S. (2005), *Causal models: How people think about the world and its alternatives*, Oxford University Press.

Stalnaker, R. (1968), A theory of conditionals, *in* 'Ifs', Springer, pp. 41–55.