Causal Models and the Logic of Counterfactuals

Jonathan Vandenburgh*

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 limitations compared to the alternative similarity-based approach: they only apply to a limited subset of counterfactuals and the connection to counterfactual logic is not straightforward. This paper argues that these limitations arise from the theory of interventions where intervening on variables requires changing structural equations rather than the values of variables. Using an alternative theory of exogenous interventions, this paper extends the causal approach to counterfactuals to handle more complex counterfactuals, including backtracking counterfactuals and those with logically complex antecedents. The theory also validates familiar principles of counterfactual logic and offers an explanation for counterfactual disagreement and backtracking readings of forward counterfactuals.

Keywords: Counterfactuals, Causal Models, Interventions, Backtracking

Recently, causal models have received increased attention for the semantics of counterfactual sentences like 'If A were the case, then C would be the case', written $A > C$.[1] Causal accounts of counterfactuals rely on the concept of intervention: a counterfactual is true if C is true when one intervenes to set A true. This approach to counterfactual semantics connects counterfactual language with other aspects of human reasoning studied with causal models (Glymour, 2001; Sloman, 2005; Gopnik and Schulz, 2007) and with empirical work on counterfactual inference.[2]

Despite the potential of causal approaches, many philosophers prefer similarity-based models of counterfactuals. Following Lewis (2013) and Stalnaker (1968), similarity-based models propose that a counterfactual $A > C$ is true if C is true in the most similar world(s) where A is true. The main advantages of similarity-based models are that they apply to a broader range of counterfactual sentences

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

[1] See the classic works of Galles and Pearl (1998) and Pearl (2009), as well as more recent work: Briggs (2012); Kaufmann (2013); Santorio (2014); Ciardelli et al. (2018).

[2] 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 developed games exclusively compatible with one platform (Lee, 2013).

and correspond nicely to counterfactual logics. Many causal models of counterfactuals, including that of Pearl (2009), cannot explain backtracking counterfactuals, such as those where the antecedent is the effect rather than the cause of the consequent ('If the grass were wet, then it must have rained'). Furthermore, most causal theories of counterfactuals only apply only to counterfactuals with antecedents which are conjunctions of variable assignments (Hiddleston, 2005; Pearl, 2009; Halpern, 2013), and the most promising extension to logically complex counterfactuals (Briggs, 2012) violates modus ponens, a standard principle of counterfactual logic.

In this paper, I argue that we can overcome these limitations by invoking a different theory of causal intervention. While Pearl, and the models following his account, argue that interventions require changing the structural equations of a causal model, I argue that we get better results for counterfactual truth conditions if interventions instead change the values of exogenous variables. In particular, I argue that a counterfactual semantics built on exogenous interventions can analyze backtracking counterfactuals, extend to logically complex antecedents, and satisfy familiar axioms of counterfactual logic, including modus ponens.

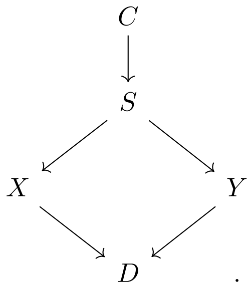
The paper is organized as follows. In §1, I introduce the foundations of causal models and the notion of an intervention, highlighting how Pearl's approach fails to predict the expected truth values for backtracking counterfactuals and motivating the notion of exogenous intervention. In §2, I define exogenous interventions more formally, characterizing the set of interventions which force a counterfactual antecedent A and motivating a minimality condition. I then use this to define a selection function for counterfactual semantics, and in §3, I show that this selection function satisfies the logical axioms for a familiar logic of counterfactuals, Pollock's (1981) counterfactual logic **SS**. In §4, I compare this account with that of Hiddleston (2005), who also offers a causal framework which can analyze backtracking counterfactuals. I note that, while his model can only handle a more limited set of counterfactuals, it motivates a stronger notion of minimality than defined in §2, showing how one can define competing counterfactual semantics within the exogenous intervention approach. In §5, I discuss the difference between exogenous interventions and Pearl's model in greater depth, showing how one can replicate many of Pearl's predictions using exogenous interventions without the logical limitations of his approach. In §6, I discuss two further aspects of counterfactual reasoning which the exogenous intervention account can explain: counterfactual disagreement and backtracking readings of forward counterfactuals.

1 Causal Models and Interventions

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 interventions and counterfactual truth conditions, we must define causal models more formally.[3] 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 is an assignment to the parents PA_i of V_i and u_i is the assignment to $U_i \subseteq U$, 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 values 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, we fix the values of all endogenous variables. If we let $\mathcal{U}$ represent the set of possible values of the exogenous variables and $\mathcal{V}$ the set of values of endogenous variables, the set of structural equations F forms a function from exogenous variable assignments to endogenous variable assignments, $F : \mathcal{U} \to \mathcal{V}$. Therefore, the values of the

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

endogenous variables in a causal model are completely determined by the structural equations and the values of the exogenous variables. 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]

Causal approaches to the truth conditions of a counterfactual $A > C$ often proceed by considering interventions in the causal model which set A true. The dominant approach to counterfactual interventions follows Pearl, who argues that interventions replace the structural equations for endogenous variables. Consider again the counterfactual 'If X were to shoot, then the prisoner would die.' On Pearl's approach, intervening to fix the antecedent replaces the structural equation $X = C$ with the structural equation $X = 1$. This intervention breaks the causal laws of the model, rendering the antecedent fixed regardless of the values of the parent variables. This is meant to capture the intuitive difference between intervention and observation: intervention involves hypothetically changing the laws of the model, while observation involves observing a realization consistent with the laws (Hagmayer et al., 2007; Fisher, 2017a).

This notion of an intervention becomes problematic, however, when one considers backtracking counterfactuals, such as counterfactuals where the consequent causes 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, this kind of reasoning is excluded by Pearl's approach to interventions: 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 $S = 1$. This is a problem for Pearl's approach to interventions, as there are many cases where backtracking appears correct in counterfactual reasoning. This is illustrated by other examples of backtracking counterfactuals, such as 'If the grass were wet, then it must have rained' or 'If the light were on, the switch would be up.' There is also experimental evidence supporting backtracking in counterfactual reasoning (Rips, 2010; Gerstenberg et al., 2013).

In backtracking reasoning, we keep the laws, or structural equations, of the causal model the same, instead considering changes to the variables in the model which make the antecedent true. This motivates an alternative conception of intervention: an intervention is a change to the values of exogenous variables in a causal model.[6] Consider how this works in the above case: C is the only

[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).

[6] This approach to interventions is also introduced in LeRoy (2019), though not for the

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 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, $u \in \mathcal{U}$. Since all endogenous variables are determined by an assignment $u \in \mathcal{U}$, elements of $\mathcal{U}$ play the role of truthmakers for propositions of variable assignments, and we can associate propositions built from variable assignments with sets of worlds. Assuming the causal model is fixed across worlds, we can simply treat the exogenous variable assignment u as the causal world.[7]

truth conditions of counterfactuals.

[7] Fixing the causal model poses problems for 'counternomic' or 'counterlegal' counterfactuals, where the counterfactual requires breaking the laws of the causal model. See Fisher (2017b).

If $V_i = v_i$ is an endogenous variable assignment, this determines a set of possible worlds by $[V_i = v_i] = \{u \in \mathcal{U} : F(u)_i = v_i\} \subseteq \mathcal{U}$, 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 $\mathcal{U}$ as 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 of 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 eight 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: $(U_C, U_X, U_Y) = (0, 1, 0)$ and $(U_C, U_X, U_Y) = (0, 0, 1)$.

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 based on an independence assumption. However, as with similarity models of counterfactual semantics, one could argue for further restrictions on the selection function; I discuss one possible such restriction motivated by Hiddleston's theory in §4.

For i to be an intervention forcing A in u, i must involve a change to exogenous variables which makes A true. However, not all such variable changes are relevant for counterfactual intervention. Interventions which change variables independent of A, for example, require intervening on the world to change more than what is necessary to realize A. Consider a counterfactual like 'If John's shirt were green, he would be the same height,' which we judge true. Here, we expect the relevant interventions to change the color of John's shirt, but not to have any effect on his height: we should not include the intervention where John's shirt becomes green and he becomes taller in the selection function. This motivates the formal characterization of interventions introduced below: A-interventions are the partial variable assignments which are minimally necessary to produce A in u.

Suppose there are m exogenous variables, so $U = (U_1, ..., U_m)$, and let $S \subseteq \{1, ..., m\}$ be a set of indices with complement S'. For any $u \in \mathcal{U}$, let $u|_S$ represent the projection of u onto the indices in S and $\mathcal{U}_S$ the set of all possible variable assignments to exogenous variables indexed by S. A partial variable assignment r_S is a variable assignment to the set of variables indexed by the set of indices S, or an element $r_S \in \mathcal{U}_S$. For a variable assignment $u|_S$ to the variables indexed by S and $u|_{S'}$ to the variables indexed by S', let $u|_S \bigoplus_S u|_{S'}$ represent the unique complete variable assignment in $\mathcal{U}$ which restricts to $u|_S$ on S and $u|_{S'}$ on S'.

We can now use a partial variable assignment r_S to intervene in a world u. For $r_S \in \mathcal{U}_S$ and $u \in \mathcal{U}$, we define the world where we intervene on u by r_S as $u|r_S = r_S \bigoplus_S u|_{S'}$. This is the world where we change the values of u on S to the values r_S, but leave all other variables unchanged. We then define the set of restricted variable assignments which force A in a world u:

$$R_u(A) = \{r_S : r_S \in \mathcal{U}_S \ \& \ u|r_S \in [A]\}.$$

This is the set of partial variable assignments such that imposing these variable assignments on the world u gives a world $u|r_S$ 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 do not 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. This means that, if i_S is a minimal intervention fixing A, one should not be able to fix A while making a smaller subset of the changes that i_S makes. Otherwise, some of the variable changes required by i_S would be independent of A in the sense that making the additional changes has no effect on the value of A. As argued above, we wish to only include those changes which are directly relevant to realizing A. This motivates defining an order $\leq$ on $R_u(A)$. Suppose $r_{S_1}, r'_{S_2} \in R_u(A)$ assign variables S_1 and S_2. We say $r_{S_1} \leq r'_{S_2}$ iff r'_{S_2} is an extension of r_{S_1}, so $S_1 \subseteq S_2$ and $r'_{S_2}|_{S_1} = r_{S_1}$. We can now define the set of interventions which force A, $I_u(A)$, as the $\leq$-minimal elements of $R_u(A)$:

$$I_u(A) = \{i_S \in R_u(A) : \nexists r_{S'} \in R_u(A), r_{S'} \neq i_S, r_{S'} \leq i_S\}.$$

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:

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

Note that this definition applies to all propositions A and C built out of variable assignments. Furthermore, since this definition now associates a set of variable assignments with a counterfactual sentence, we can ascribe truth values to right-nested counterfactuals, where A is a non-counterfactual proposition and C contains counterfactual terms without counterfactual antecedents.[8]

To see how these truth conditions work for both forward and backtracking counterfactuals, 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 $U_Y = 1$, $S = 0$, so the counterfactual is false. This makes sense given the model: 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. Note, however, that if one did not include the possibility of X and Y deviating from the signal through the exogenous variables U_X and U_Y, this backtracking counterfactual would be true, as in the model from the previous section.

3 Logic of Exogenous Intervention Models

Similarity-based models for counterfactuals rely on selection functions $f(A, u) : \mathcal{P}(\mathcal{U}) \times \mathcal{U} \to \mathcal{P}(\mathcal{U})$, 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 $f(A, u) = \{u|i_S : i_S \in I_u(A)\}$. 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 $>$.[9] 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 $f(A, u) = \{u\}$
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: $u \in [A > C]$ 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′**:

[8] An example of a right-nested counterfactual will be discussed in §5. Note that left-nested counterfactuals are often excluded from counterfactual analysis, c.f. Briggs (2012); this issue is also discussed in Vandenburgh (2020).

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

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 $w = u|i_S$ for some $i_S \in I_u(A)$. Since $i_S \in R_u(A)$, $u|i_S \in [A]$ 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_0, which changes no exogenous variables, is in $R_u(A)$ since $u|i_0 = u \in R_u(A)$. Since $i_0 \leq r_S$ for every other possible intervention $r_S \in R_u(A)$, i_0 is the unique minimal element in $R_u(A)$ and the only element in $I_u(A)$. Since $f(A, u) = \{u|i_S : i_S \in I_u(A)\}$, $f(A, u) = \{u|i_0\} = \{u\}$. □

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 $i_S \in I_u(A)$, there is some $j_{S^*} \in I_u(B)$ such that $u|i_S = u|j_{S^*}$. Suppose $i_S \in I_u(A)$. Since $f(A, u) \subseteq [B]$, $u|i_S \in [B]$, so $i_S \in R_u(B)$. Then there is a $j_{S^*} \in I_u(B)$ such that i_S extends j_{S^*}. But since $j_{S^*} \in I_u(B)$ and $f(B, u) \subseteq [A]$, $u|j_{S^*} \in [A]$, so $j_{S^*} \in R_u(A)$. This means there is an $i'_{S'} \in I_u(A)$ such that j_{S^*} extends $i'_{S'}$. But since i_S and $i'_{S'}$ are both $\leq$-minimal elements and $i'_{S'} \leq j_{S^*} \leq i_S$, $i_S = i'_{S'} = j_{S^*}$, so $u|i_S = u|j_{S^*}$. Since we have shown $\forall i_S \in I_u(A), \exists j_{S^*} \in I_u(B)$ such that $u|i_S = u|j_{S^*}$, 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 $u|i_S \in f(A \vee B, u)$, where $i_S \in I_u(A \vee B)$. Since $u|i_S \in [A \vee B]$ by **CS1**, $u|i_S \in [A]$ or $u|i_S \in [B]$. Suppose $u|i_S \in [A]$. Then $i_S \in R_u(A)$, so there is some $j_{S^*} \in I_u(A)$ such that i_S extends j_{S^*}. Since $j_{S^*} \in I_u(A)$, $u|j_{S^*} \in [A] \subseteq [A \vee B]$, so $j_{S^*} \in R_u(A \vee B)$. This means there is some $i'_{S'} \in I_u(A \vee B)$ such that j_{S^*} extends $i'_{S'}$. But since $i'_{S'} \leq j_{S^*} \leq i_S$ and i_S and $i'_{S'}$ are both $\leq$-minimal, $i_S = i'_{S'} = j_{S^*}$, so $\exists j_{S^*} \in I_u(A)$ such that $u|i_S = u|j_{S^*}$, so $u|i_S \in f(A, u) \cup f(B, u)$. If $u|i_S \in [B]$, a parallel proof shows that $u|i_S \in f(B, u) \subseteq f(A, u) \cup f(B, u)$. Therefore, $f(A \vee B, u) \subseteq f(A, u) \cup f(B, u)$. □

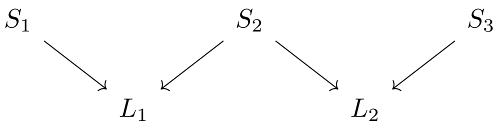
CS6: $u \in [A > C]$ iff $f(A, u) \subseteq [C]$

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.

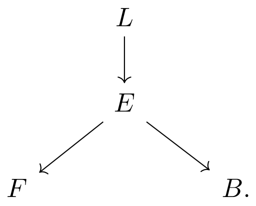
This shows that the exogenous intervention approach to counterfactuals, in addition to handling both forward and backtracking counterfactuals, has familiar logical properties. Counterfactual models built on Pearl's approach to interventions either do not extend to logically complex antecedents, such as disjunctive antecedents, (Galles and Pearl, 1998; Halpern, 2000) or require abandoning familiar logical principles such as modus ponens, a consequence of strong centering, **CS2** (Briggs, 2012). Note that the exogenous intervention approach is consistent with further possible restrictions on the selection function: as discussed in the next section, Hiddleston offers a stronger conception of a 'minimal break' which leads to slightly different predictions for the truth conditions of counterfactuals. The classification of specific semantics built on the exogenous intervention approach through completeness results is left for future work.

4 Comparison to Hiddleston's Model

Hiddleston's model of the truth conditions of counterfactuals is most similar to the exogenous intervention model. 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, where the 'breaks' considered for a model change the values of variables rather than the structural equations. However, Hiddleston's model differs both formally and substantively from the exogenous intervention model. Formally, Hiddleston does not differentiate exogenous and endogenous variables and permits indeterministic structural equations. Substantively, Hiddleston uses a stronger notion of minimality than that motivated above. Furthermore, Hiddleston's model does not apply to logically complex antecedents and has not been proven to satisfy axioms of conditional logic. After discussing the differences between the formal elements of the two theories, I will introduce Hiddleston's stronger notion of a minimal break and show how one could incorporate such a stronger notion in the exogenous intervention model. This motivates both the flexibility of the model and the benefits of future work comparing the logical and semantic properties of minimality conditions on exogenous interventions.

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 $\Pr(Y = y|X = x) = p$. The inclusion of exogenous variables is a minor difference: for a model with only endogenous variables, any variable which has no parents in the graph $\mathcal{G}$ over V can be freely set with no constraints from its structural equation, since structural equations only involve the parents of a variable. This means that any such variable can be treated as exogenous: 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 $V_i = U_i$.

Hiddleston's use of indeterministic structural equations, on the other hand, introduces a more significant difference from the framework here. However, as I will argue, the notion of error variables will allow such examples to be reformulated in the exogenous intervention model. To see how indeterministic structural equations work, 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, in Hiddleston's theory, are $\Pr(E = 1|L = 1) = 0.95$, $\Pr(E = 1|L = 0) = 0$, $F = E$, and $B = E$ with causal graph:



We can handle such indeterminacies with 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 with $E = L \wedge \neg U'_E$, 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

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

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.

Noting these two differences allows us to translate Hiddleston's examples into the framework of causal models used in §2 and therefore compare the predictions of the two models for counterfactual sentences. For example, to translate the case 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 and write the new structural equations $L = U_L$, $E = L \wedge \neg U'_E$, $F = E$, and $B = E$. Analyzing this case also indicates the benefit of replacing indeterministic structural equations with exogenous error variables. 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.' Even though, most of the time, an explosion would happen, Hiddleston predicts this counterfactual is false: there is always a possible outcome where the cannon is lit but an explosion does not occur. However, in the exogenous intervention model, this counterfactual is true in 95% of worlds (where $U'_E = 0$) and false in the other 5%. This, I argue, is a more reasonable approach to the truth conditions of counterfactuals, as there are many cases where counterfactuals are generally considered true, even with a small probability of error.

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 $\mathcal{V}(u|i)$ 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 $\mathcal{V}(u|i) \cap Z$ the largest subset of variables from Z which have the same value in $\mathcal{V}(u|i)$ and in u. We can then define an order on $I_u(A)$, $\leq_H$, by saying $i \leq_H i'$ iff $\mathcal{V}(u|i') \cap Z \subseteq \mathcal{V}(u|i) \cap Z$.

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, $(U_L, U'_E) = (1, 0)$. Consider the counterfactual 'If the flash hadn't occurred, the cannon was still lit,' $\neg F > L$. 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 model, we can recover a restricted counterfactual semantics with slightly different truth conditions than found in §2. This suggests, more generally, that there is value in considering possible restrictions to the framework introduced in §2, which can result in slightly different semantic and logical predictions for counterfactuals.

5 Comparison to Endogenous Interventions: Pearl and Briggs

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. And he has good reason for this conclusion: backtracking in counterfactual reasoning 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 introducing additional exogenous variables. 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 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 an exogenous variable 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.

Including sufficiently many exogenous variables in the model allows us to replicate many of the predictions which arise from Pearl's account of interventions and counterfactual truth conditions. For any endogenous variable in a causal model which does not include the possibility of an exogenous shock, one could simply add 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. Setting the exogenous variable $U_i = u_i$ then corresponds to the intervention where one replaces the structural equation for V_i by $V_i = u_i$. This interpretation in terms of exogenous interventions satisfies a common characterization of interventions in causal models (Hagmayer et al., 2007; Fisher, 2017a), where intervening on

[11] One way of measuring this in the US is by noting when the Fed funds rate deviates from futures on the Fed funds rate. See Kuttner (2001).

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

V_i forces V_i to be independent of its parents: activating U_i so that $U_i \neq$ OFF leaves V_i independent of its parents.

One major point of difference, which I take to be an advantage for the exogenous intervention model, is that the exogenous intervention model leads to a theory of counterfactuals which satisfies strong center (**CS2**), and therefore modus ponens, while Pearl's approach does not. When evaluating a counterfactual where the antecedent A is actual, Pearl's approach requires changing the structural equations of the model, affecting some aspects of the actual world, while the exogenous intervention model predicts that no intervention is required to set the antecedent true. This is illustrated by the extension of Pearl's approach to logically complex counterfactuals in Briggs (2012), who notes that counterexamples to modus ponens arise in Pearl's approach for right-nested counterfactuals.

Consider the modified firing squad case where X and Y can shoot independently, without signal S. While one could add additional exogenous variables to imitate Pearl's predictions in more cases, I will focus on counterfactuals where the model with the three exogenous variables U_C, U_X, and U_Y is sufficiently rich. Consider the nested counterfactual 'If X were to shoot, then if the court had not ordered it, the prisoner would die,' represented $X > (\neg C > D)$.[13] Assume the world is $(U_C, U_X, U_Y) = (1, 0, 0)$. Since $U_C = 1$, $X = 1$, so X is true. On Pearl's semantics, we evaluate $X > (\neg C > D)$ by replacing the structural equation $X = S \vee U_X$ with $X = 1$ and then evaluating $\neg C > D$. We evaluate this by replacing $C = U_C$ with $C = 0$, but since the new structural equation tells us $X = 1$, we still have $D = 1$, so this counterfactual is true, meaning $X > (\neg C > D)$ is true. However, $\neg C > D$ itself is false: setting $C = 0$ without antecedently setting $X = 1$ entails that $D = 0$, so the prisoner does not die. This violates modus ponens: X and $X > (\neg C > D)$ are both true, but $\neg C > D$ is false.

Contrast this with the corresponding interpretation in the exogenous intervention model. Here, X is still true, and since intervening to set $\neg C$ corresponds to setting $U_C = 0$, which leads to $D = 0$, $\neg C > D$ is also false. However, unlike on Pearl's approach, the counterfactual $X > (\neg C > D)$ is also false: since X is true in the world, no intervention on exogenous variables is necessary to set X true, so the consequent $\neg C > D$ is simply evaluated in the actual world, where it is false. This verdict arises because the exogenous intervention model satisfies strong centering, where if the antecedent is true in the actual world u, u is the unique relevant world for counterfactual evaluation. This discussion shows that the exogenous intervention model can incorporate many of the insights of Pearl-style interventions while providing more intuitive verdicts for logically complex counterfactuals.

[13] Adapted from Briggs (2012, p. 150).

6 Ambiguity in Counterfactual Semantics

Throughout, I have emphasized how using exogenous interventions to understand the truth conditions of counterfactuals can explain both forward and backtracking readings of counterfactuals. When a causal model includes a rich set of exogenous variables, as in models which aim to replicate Pearl's approach to interventions, backtracking readings of counterfactuals are suppressed. On the other hand, when structural equations include fewer possibilities for error, counterfactuals leave causal relationships intact through backtracking readings. We saw this in the case of monetary policy intervention, as well as in the original firing squad case, where the truth value of the counterfactual 'If X shoots, then the captain gave the signal' depends on whether we include an exogenous variable governing Shooter X's ability to shoot without a signal.

While one might think that the dependence of truth conditions on model variables introduces too much context sensitivity, there is good reason to think that such flexibility is necessary to accurately represent counterfactual reasoning. There are many cases, such as the case of monetary policy intervention, where agents (in this case, economists) disagree explicitly about the causal model and, specifically, which variables are and are not subject to exogenous shocks. In these cases, where people, even experts, do not agree on the correct causal model, we expect disagreement on the truth values of counterfactual sentences.[14]

An instance of this kind of disagreement which has received attention in the philosophical literature is the possibility of both forward and backtracking readings of the same counterfactual. 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 in her model of the world, even though it was true in 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

[14] Note that I leave open what constraints, if any, govern the choice of an appropriate causal model. For one approach to this problem, see Woodward (2016).

the counterfactuals.[15]

Note that, among theories of counterfactual interventions, the exogenous intervention model is uniquely suited to explain how forward and backtracking readings can arise for the same counterfactual. Pearl's approach to interventions does not allow for backtracking readings, offering no way of explaining backtracking readings of forward counterfactuals. While Hiddleston's theory may be able to incorporate both readings, the account would be less clear without exogenous variables. In the above case, Hiddleston would need to either include an endogenous variable specifying the condition which prevents Smith from dying or replace the deterministic structural equation with an indeterministic one, though without the ability to comment on which worlds validate the forward counterfactual and which worlds validate the backtracking counterfactual. The exogenous intervention model counters Lee's (2015) claim that the ambiguity of counterfactuals exhibited in forward and backtracking readings requires separate causal models for intervention and extrapolation, offering an explanation for how both readings can arise with one theory of intervention.[16]

7 Conclusion

In this paper, I argued for the use of exogenous interventions to capture the semantics of counterfactual sentences. On this approach, a counterfactual $A > C$ is true in a causal world u if C is true in all worlds formed by intervening to set A true, where an intervention is a change to exogenous variables rather than structural equations. In contrast to competing models, this approach can handle both forward and backtracking counterfactuals, applies to logically complex antecedents, and satisfies the axioms of a familiar counterfactual logic, Pollock's **SS**. This approach can be extended by considering additional restrictions on the selection function, as illustrated in the reformulation of Hiddleston's theory in §4, and can capture many of the intuitions of Pearl's approach to counterfactuals, provided the model includes sufficiently many exogenous variables. The sensitivity of the framework to the choice of exogenous variables can also explain disagreement about counterfactuals and how we get forward and backtracking interpretations of the same counterfactuals, a fact other causal approaches to counterfactuals fail to explain.

[15] 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).

[16] 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 §5 where we associate interventions with exogenous variables resolves this division between intervention and extrapolation.

References

Arlo-Costa, H. (2019). "The logic of conditionals." *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, ed., (https://plato.stanford.edu/archives/sum2019/entries/logic-conditionals/), 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. (2017a). "Causal counterfactuals are not interventionist counterfactuals." *Synthese*, 194(12), 4935–4957.

Fisher, T. (2017b). "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." *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 35, 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.

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., and Waldmann, M. R. (2007). "Causal reasoning through intervention." *Causal learning: Psychology, philosophy, and computation*, 86–100.

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.” *International Workshop on Logic, Rationality and Interaction*, Springer, 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.

LeRoy, S. F. (2019). “Causal inference.” *UC Santa Barbara: Department of Economics*.

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

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

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

Pollock, J. L. (1981). “A refined theory of counterfactuals.” *Journal of Philosophical Logic*, 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.” *Semantics and Linguistic Theory*, Vol. 24, 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.” *Ifs*, Springer, 41–55.

Vandenburgh, J. (2020). "Conditional learning through causal models." *Synthese* https://doi.org/10.1007/s11229-020-02891-x.

Woodward, J. (2016). "The problem of variable choice." *Synthese*, 193(4), 1047–1072.