Backtracking through interventions: An exogenous intervention model for counterfactual semantics

Jonathan Vandenburgh

November 21, 2022

Note: This is a pre-copyedit version of an article published in *Mind & Language*. Please refer to the final version: https://doi.org/10.1111/mila.12447.

Abstract

Causal models show promise as a foundation for the semantics of counterfactual sentences. However, current approaches face limitations compared to the alternative similarity theory: they only apply to a limited subset of counterfactuals and the connection to counterfactual logic is not straightforward. This paper addresses these difficulties using exogenous interventions, where causal interventions change the values of exogenous variables rather than structural equations. This model accommodates judgments about backtracking counterfactuals, extends to logically complex counterfactuals, and validates familiar principles of counterfactual logic. This combines the interventionist intuitions of the causal approach with the logical advantages of the similarity approach.

Keywords: counterfactuals, causality, interventions, backtracking, structural equations

1 Introduction

Consider the counterfactual Fine (1975) raises against the similarity theory of counterfactuals: "If Nixon had pressed the button, there would have been a nuclear holocaust." Intuitively, this counterfactual is true, as Nixon pressing the button would have caused a nuclear holocaust. This judgment, however, conflicts with the most natural interpretation of the similarity theory of counterfactuals. On this theory, following Stalnaker (1968) and Lewis (2013), a counterfactual "If A had been the case, then C would have been the case," written $A > C$, is true if C is true in the closest or most similar world(s) where A is true. Since a world where a nuclear holocaust occurs is intuitively less similar to our world than a world where something intervenes to prevent a nuclear holocaust, the similarity theory of counterfactuals seems to incorrectly predict that this counterfactual is false.

Causal theories of counterfactuals can more easily explain judgments in cases which conflict with intuitions about similarity.[1] The causal account relies on the concept of intervention: a counterfactual A $>$ C is true if C is true when one intervenes to set A true.[2] For example, if a causal intervention had forced Nixon to press the button, then a nuclear holocaust would have resulted, regardless of how distant or dissimilar this "intervened world" is from the actual world. Causal theories of counterfactuals also connect counterfactual language with other aspects of human reasoning studied with causal models (Glymour, 2001; Sloman, 2005; Gopnik & Schulz, 2007) and with empirical work on counterfactual inference.[3]

Despite the potential of causal theories of counterfactuals, many philosophers prefer the similarity theory. Some reasons for this include that the similarity theory applies to a broader range of counterfactual sentences and corresponds nicely to counterfactual logics. One significant limitation of causal theories of counterfactuals, particularly theories following Pearl (2009), is that they cannot explain backtracking counterfactuals. These are counterfactuals where the antecedent is the effect rather than the cause of the consequent, such as "If the microwave had been on, it would have been plugged in." Furthermore, most causal theories of counterfactuals only apply to a logically restricted class of counterfactuals, excluding counterfactuals with disjunctive antecedents (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 most authors, following Pearl, 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 validate familiar properties of counterfactual logic, including modus ponens. Thus, the exogenous intervention model can capture the intuitive appeal of causal approaches to counterfactual semantics while retaining the logical advantages which come with theories built on similarity.

The paper is organized as follows. In Section 2, I introduce the formalism of causal models and the notion of an intervention, highlighting how Pearl's approach fails to predict the expected truth values for backtracking counterfac-

[1] Even Lewis's (1979) account of Fine's example within the similarity theory invokes laws, a central component of causal theories.

[2] See the classic works of Galles and Pearl (1998) and Pearl (2009), as well as more recent work: Briggs (2012); Kaufmann (2013); Ciardelli et al. (2018); Santorio (2019).

[3] 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).

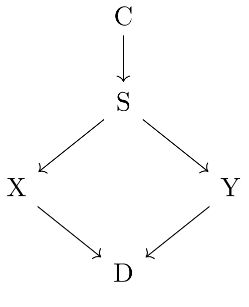
tuals and motivating exogenous interventions. In Section 3, I define exogenous interventions more formally, characterizing the set of interventions which force a counterfactual antecedent. I then use this to define a selection function for counterfactual semantics which satisfies the axioms for a familiar logic of counterfactuals, Pollock's (1981) counterfactual logic SS, as demonstrated in the Appendix. In Section 4, I discuss how one can use the exogenous intervention framework to represent more restrictive semantic theories of counterfactuals, like those based on similarity or those implementing a further minimality condition found in Hiddleston (2005). In Section 5, I discuss the differences 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.

2 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), the 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, which represent the causal relationships in the model, 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. Suppose that the court did not order execution and that the prisoner lived, and consider the counterfactual “If X had shot, the prisoner would have died.” If we make an intervention on the causal model to set $\mathrm{X} = 1$, so that Shooter X shot, then since $\mathrm{D} = \mathrm{X} \vee \mathrm{Y}$, $\mathrm{D} = 1$, so the prisoner would have died; 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.[4] A causal model $\mathcal{M} = (\mathrm{U}, \mathrm{V}, f_i)$ consists of a finite set of exogenous variables U, a finite set of endogenous variables $\mathrm{V} = (\mathrm{V}_1, ..., \mathrm{V}_n)$, and a set of structural equations $\mathrm{F} = (f_1, ..., f_n)$, 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 $\mathrm{U}_i \subseteq \mathrm{U}$, 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, assigning values to all exogenous variables is sufficient to determine the values of all endogenous variables in the model. Thus, 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, $\mathrm{F} : \mathcal{U} \rightarrow \mathcal{V}$. Even though exogenous variable assignments are crucial for understanding the different configurations consistent with a causal model, the significance of exogenous variables is often overlooked in causal theories of counterfactuals. 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.[5]

The dominant causal approach to counterfactuals, following Pearl, proposes that a counterfactual $\mathrm{A} > \mathrm{C}$ is true when C is true after one has intervened to set A true, where an intervention replaces the structural equations of the original causal model to require that A is true. Consider again the counterfactual “If X had shot, then the prisoner would have died.” On Pearl’s approach, intervening to fix the antecedent replaces the structural equation $\mathrm{X} = \mathrm{S}$ with the structural equation $\mathrm{X} = 1$. 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 approach to interventions, however, is limited in the kinds of counterfactuals it can address. In particular, intervening by changing structural equations cannot explain judgments for backtracking counterfactuals where the consequent causes the antecedent. Consider the counterfactual “If X had shot,

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

[5] 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 $\mathrm{U_C}$ such that $\mathrm{C} = \mathrm{U_C}$.

the captain would have signaled for it." Intuitively, this counterfactual is true: if the causal model is correct, X only shoots if the captain signaled to, so X = 1 only if S = 1. This judgment involves reasoning backwards from effect to cause: for X = 1 to have been true, S = 1 must have caused X = 1, since S = 1 is the only possible cause of X = 1 in the model. The philosophical literature provides further examples of backtracking counterfactuals (Khoo, 2017), and experimental evidence suggests that many people accept backtracking readings of counterfactuals (Rips, 2010; Gerstenberg et al., 2013). Backtracking counterfactuals are especially compelling in cases where all possible causes are accounted for, like "If the light had been on, the light switch would have been up" or "If the microwave had been on, it would have been plugged in."

Pearl's approach to interventions, however, cannot explain backtracking judgments: intervening on A by changing the structural equations for variables in A can never change the factors upstream from A. For example, intervening so that X shot changes the structural equation for X from X = S to X = 1, leaving all variables upstream unchanged. Under this intervention, the captain would still not have signaled and the court would still have not ordered execution, even though these changes are necessary for X to have shot. This applies to other backtracking counterfactuals as well: on Pearl's theory, intervening to turn a microwave on does not require that the microwave be plugged in, so Pearl's theory offers no way to account for the truth of the counterfactual "If the microwave had been on, it would have been plugged in."

Backtracking reasoning involves keeping the laws, or structural equations, of the causal model the same and instead reasoning about the variables in the model which would have to change to make the antecedent true. Since all endogenous variable values are determined by the exogenous variables, the only way to change the values of endogenous variables without changing the structural equations is to change the values of the exogenous variables. This motivates an alternative conception of intervention: an intervention is a change to the values of exogenous variables in a causal model.[6] For example, in the firing squad case, 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 had shot, the captain would have signaled for it": intervening to set X = 1 involves setting C = 1, which sets S = 1, so the counterfactual is always true. Exogenous interventions can handle both forward and backtracking counterfactuals, challenging arguments that forward and backtracking reasoning arise from different causal procedures, such as intervention and extrapolation (Lucas & Kemp, 2012, 2015; Lee, 2015).

With exogenous interventions, counterfactual truth conditions depend on the

[6] This approach to interventions is also introduced in LeRoy (2020), though not for the truth conditions of counterfactuals.

exogenous variables which are included in the model.[7] Suppose, for example, that we think it is more accurate to attribute to X the possibility of shooting without receiving the signal. To account for this possibility, we should add an exogenous variable $\mathrm{U_X}$ to the causal model such that $\mathrm{X = S \vee U_X}$, even if we consider the activation of $\mathrm{U_X}$ extremely unlikely. Exogenous variables like $\mathrm{U_X}$ are sometimes referred to as error terms because they introduce the possibility of outcomes deviating from the expected course of events. With this additional exogenous variable, intervening to make X shoot ($\mathrm{X} = 1$) no longer requires that the captain gave the signal ($\mathrm{S} = 1$) or that the court ordered execution ($\mathrm{C} = 1$), as the exogenous intervention $\mathrm{U_X} = 1$ can cause X to shoot without the signal to do so. Thus, while "If X had shot, the captain would have signaled for it" is true in the original model, it need not be true in a model where X can choose to shoot exogenously.[8]

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

3 The Exogenous Intervention Model

Before discussing how exogenous interventions can serve as the foundation for counterfactual truth conditions, I will introduce the notion of a causal world, allowing for the use of tools from the similarity theory of counterfactuals. Pearl (2009) defines causal worlds, but makes little use of the notion in his analysis, and the notion is largely left out of later causal theories of 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

[7] Determining the correct causal model for a situation is a challenging issue (Halpern & Hitchcock, 2010; Woodward, 2016). While I will not address this issue directly, some concerns which may be relevant for whether to include an exogenous variable are: the probability or frequency of activation of an exogenous variable, whether the exogenous variable is activated in the actual world, and whether the exogenous variable is made contextually salient.

[8] The fact that the exogenous intervention model can accommodate both forward and backtracking readings of counterfactuals may also be useful for understanding backtracking readings of forward counterfactuals. Suppose your friend Smith is on top of a building about to jump, but steps off (Jackson, 1977; Khoo, 2017). On an ordinary forward reading, the counterfactual "If Smith had jumped, he would have died" is true. Your friend Beth, however, thinks 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 have died" is false, offering a backtracking reading of the same counterfactual. Both judgments are consistent with the exogenous intervention model, with different judgments resulting from different causal models. Beth's backtracking reasoning proposes that there is an inhibiting abnormality preventing Smith from dying if he jumps, making the counterfactual false since intervening to make Smith jump no longer leads to his death. On the usual forward reading, there is no such inhibiting abnormality activated in the actual world, so the counterfactual is true.

worlds, we can simply treat the exogenous variable assignment u as the causal world.[9]

If $\mathrm{V}_i = v_i$ is an endogenous variable assignment, this determines a set of possible worlds by $[\mathrm{V}_i = v_i] = \{u \in \mathcal{U} : \mathrm{F}(u)_i = v_i\} \subseteq \mathcal{U}$, so $u \in [\mathrm{V}_i = v_i]$ iff $\mathrm{V}_i = v_i$ 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, where negation, conjunction, and disjunction correspond to set-theoretic complementation, intersection, and union, respectively. We refer to subsets of $\mathcal{U}$, or sets of causal worlds, as propositions. Counterfactual truth conditions will be defined for all propositions, allowing the exogenous intervention model to incorporate counterfactuals built from logically complex propositions.

For illustration, consider a modified version of the firing squad example where both X and Y are able to shoot without receiving the signal. Here, we have the exogenous variable $\mathrm{U_C}$ representing whether the court orders execution as well as exogenous variables $\mathrm{U_X}$ and $\mathrm{U_Y}$ representing the decisions of X and Y to shoot regardless of signal. The structural equations for X and Y shooting are $\mathrm{X} = \mathrm{S} \vee \mathrm{U_X}$ and $\mathrm{Y} = \mathrm{S} \vee \mathrm{U_Y}$: the shooter shoots iff the captain signals for it $(\mathrm{S} = 1)$ or the shooter makes the decision to shoot regardless of the signal $(\mathrm{U_X} = 1$ or $\mathrm{U_Y} = 1)$. Thus, in the modified firing squad example, there are three exogenous variables, $\mathrm{U_C}$, $\mathrm{U_X}$, and $\mathrm{U_Y}$, with causal graph as above and structural equations $\mathrm{C} = \mathrm{U_C}$, $\mathrm{S} = \mathrm{C}$, $\mathrm{X} = \mathrm{S} \vee \mathrm{U_X}$, $\mathrm{Y} = \mathrm{S} \vee \mathrm{U_Y}$, and $\mathrm{D} = \mathrm{X} \vee \mathrm{Y}$. In this model, there are eight possible worlds corresponding to the eight possible assignments to the three exogenous variables. Propositions in the model correspond to sets of possible worlds: "The prisoner dies and either Shooter X or Shooter Y does not shoot," for example, is true in exactly two worlds: $(\mathrm{U_C}, \mathrm{U_X}, \mathrm{U_Y}) = (0, 1, 0)$ and $(\mathrm{U_C}, \mathrm{U_X}, \mathrm{U_Y}) = (0, 0, 1)$.

Causal worlds allow for a convenient interpretation of exogenous interventions. Interventions are manipulations to a causal model to set a given proposition A true. In the exogenous intervention model, these manipulations are variable changes: given a causal world u, we can manipulate the world so that A is true by changing some of the exogenous variable values in u. However, not all variable changes which make A true arise from direct manipulation of the world to force A. Intuitively, when we think of an intervention in the world to, say, make John's shirt green, we do not think of changes to the world which also make John taller. The additional variable change making John taller is irrelevant for the proposition we are intervening to set true, and therefore should not be considered part of a causal intervention. This motivates the definition of an A-intervention, which will be given more formally below: the A-interventions in a world u are the variable changes which make A true without changing any variables unnecessary for the truth of A. A similar condition restricting interventions on A to changes necessary to make A true appears in Pearl's theory of endogenous interventions on structural equations, where causal interventions

[9] 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).

can only change structural equations for variables in A, leaving the equations for variables independent of A unchanged. However, some differences arise between these two conditions for causal intervention, as will be discussed in Section 5.

An intervention setting A true in u corresponds to changes to some of the exogenous variables in u; making these changes to u results in an "intervened world" in $\mathcal{U}$. This defines a selection function $f(\mathrm{A}, u)$ consisting of all worlds where we intervened on u to set A true. As in similarity accounts of counterfactuals, for $\mathrm{A} > \mathrm{C}$ to be true in u, C must be true in all worlds in $f(\mathrm{A}, u)$. However, since $f(\mathrm{A}, u)$ is the set of all worlds where we have intervened to make A true, this approach to counterfactuals preserves the interventionist intuition of causal theories: $\mathrm{A} > \mathrm{C}$ is true if C is true when we intervene to set A true. The selection function is built solely from the causal notion of an exogenous intervention and does not introduce any similarity-based requirements into the counterfactual semantics. However, since the exogenous intervention model shares a formal structure with the similarity theory, one could consider further restrictions on this selection function to focus only on, say, interventions which change fewer variables or which do not change variable values too much. Possible restrictions to the exogenous intervention model, including a restriction motivated by Hiddleston's (2005) theory, will be discussed in Section 4.

I will now define the set of A-interventions and the selection function for the exogenous intervention model more formally. Suppose there are m exogenous variables, so $\mathrm{U} = (\mathrm{U}_1, ..., \mathrm{U}_m)$, and let $\mathrm{S} \subseteq \{1, ..., m\}$ be a set of indices with complement $\overline{\mathrm{S}}$. For any $u \in \mathcal{U}$, let $u|_{\mathrm{S}}$ represent the projection of u onto the indices in S and $\mathcal{U}_{\mathrm{S}}$ 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 variables indexed by S, or an element $r_{\mathrm{S}} \in \mathcal{U}_{\mathrm{S}}$. For a variable assignment r_{S} to the variables indexed by S and $q_{\overline{\mathrm{S}}}$ to the variables indexed by $\overline{\mathrm{S}}$, let $r_{\mathrm{S}} \bigoplus_{\mathrm{S}} q_{\overline{\mathrm{S}}}$ represent the unique complete variable assignment in $\mathcal{U}$ which restricts to r_{S} on S and $q_{\overline{\mathrm{S}}}$ on $\overline{\mathrm{S}}$.

We can then use a partial variable assignment r_{S} to change variable values in a world u. For $r_{\mathrm{S}} \in \mathcal{U}_{\mathrm{S}}$ and $u \in \mathcal{U}$, changing variables S to r_{S} in u results in the manipulated world $u|r_{\mathrm{S}} = r_{\mathrm{S}} \bigoplus_{\mathrm{S}} u|_{\overline{\mathrm{S}}}$. 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 can then define the set of restricted variable assignments which make A true in a world u:

$$\mathrm{R}_u(\mathrm{A}) = \{r_{\mathrm{S}} : r_{\mathrm{S}} \in \mathcal{U}_{\mathrm{S}} \ \&\ u|r_{\mathrm{S}} \in [\mathrm{A}]\}.$$

This is the set of partial variable assignments such that imposing these variable assignments on the world u gives a world $u|r_{\mathrm{S}}$ where A is true. As long as a proposition A is possible, or has some world $w \in [\mathrm{A}]$ making it true, $\mathrm{R}_u(\mathrm{A}) \neq \emptyset$ since $w \in \mathrm{R}_u(\mathrm{A})$ with $\mathrm{S} = \{1, ..., m\}$; every element $w \in [\mathrm{A}]$ is in $\mathrm{R}_u(\mathrm{A})$ for any u. However, as motivated above, we do not want all elements of [A] to be A-interventions, so we must restrict attention to only the variable changes necessary to bring about A.

We want the set of A-interventions in u to be the set of variable changes in $\mathrm{R}_u(\mathrm{A})$ which bring about A without changing anything more than what is necessary to make A true. This means that, if i_{S} is an 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 unnecessary to make A true, and, as argued above, we wish to only include those changes which are directly relevant to realizing A. We can accomplish this restriction by defining an order $\leq$ on $\mathrm{R}_u(\mathrm{A})$. Suppose $r_{\mathrm{S}_1}, r'_{\mathrm{S}_2} \in \mathrm{R}_u(\mathrm{A})$ assign variables S_1 and S_2. We say $r_{\mathrm{S}_1} \leq r'_{\mathrm{S}_2}$ iff r'_{S_2} is an extension of r_{S_1}, or iff $\mathrm{S}_1 \subseteq \mathrm{S}_2$ and $r'_{\mathrm{S}_2}|_{\mathrm{S}_1} = r_{\mathrm{S}_1}$. We can now define the set of interventions which force A, $\mathrm{I}_u(\mathrm{A})$, as the $\leq$-minimal elements of $\mathrm{R}_u(\mathrm{A})$:

$$\mathrm{I}_u(\mathrm{A}) = \{i_{\mathrm{S}} \in \mathrm{R}_u(\mathrm{A}) : \nexists r_{\mathrm{S}'} \in \mathrm{R}_u(\mathrm{A}), r_{\mathrm{S}'} \neq i_{\mathrm{S}}, r_{\mathrm{S}'} \leq i_{\mathrm{S}}\}.$$

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

$$[\mathrm{A} > \mathrm{C}] = \{u \in \mathcal{U} : \forall i_{\mathrm{S}} \in \mathrm{I}_u(\mathrm{A}), u|i_{\mathrm{S}} \in [\mathrm{C}]\}.$$

This can be interpreted as a selection function semantics by taking $f(\mathrm{A}, u) = \{u|i_{\mathrm{S}}\}$ for all $i_{\mathrm{S}} \in \mathrm{I}_u(\mathrm{A})$. Note that these truth conditions apply to all propositions A and C built out of variable assignments, including disjunctions of variable assignments and propositions built from complex combinations of logical connectives. Furthermore, since we can now associate a counterfactual $\mathrm{A} > \mathrm{C}$ with the set of worlds where the counterfactual is true, $[\mathrm{A} > \mathrm{C}]$, we can ascribe truth values to right-nested counterfactuals $\mathrm{A} > \mathrm{C}$, where A is a non-counterfactual proposition and C contains counterfactual terms without counterfactual antecedents.[10] This model also leads to familiar logical properties for counterfactuals: as demonstrated in the Appendix, the selection function satisfies the axioms of Pollock's (1981) logic SS, including strong centering, which entails that modus ponens applies to counterfactuals.

To see how these definitions work, recall the modified firing squad example from above with exogenous variables $\mathrm{U_C}$, $\mathrm{U_X}$, and $\mathrm{U_Y}$ and structural equations $\mathrm{C} = \mathrm{U_C}$, $\mathrm{S} = \mathrm{C}$, $\mathrm{X} = \mathrm{S} \vee \mathrm{U_X}$, $\mathrm{Y} = \mathrm{S} \vee \mathrm{U_Y}$, and $\mathrm{D} = \mathrm{X} \vee \mathrm{Y}$. Suppose that, in the actual world, the court does not order execution and neither X nor Y choose to shoot, so $(\mathrm{U_C}, \mathrm{U_X}, \mathrm{U_Y}) = (0, 0, 0)$. Consider the counterfactual "If X had shot, the prisoner would have died." This is true if any intervention making X shoot ensures that the prisoner dies. There are two exogenous interventions forcing $\mathrm{X} = 1$: one can intervene so that the court orders execution ($\mathrm{U_C} = 1$), ensuring that the captain signals ($\mathrm{S} = 1$) and that X shoots, or one can intervene so that X chooses to shoot regardless of the signal ($\mathrm{U_X} = 1$). Note that, while other variable changes also set $\mathrm{X} = 1$, they also make unnecessary changes: any change to $\mathrm{U_Y}$ is unnecessary to set $\mathrm{X} = 1$, and once either intervention $\mathrm{U_X} = 1$ or $\mathrm{U_C} = 1$ is made, there is no need to consider a second variable change $(\mathrm{U_C}, \mathrm{U_X}) = (1, 1)$ to set $\mathrm{X} = 1$ true. Thus, $\mathrm{U_X} = 1$ and $\mathrm{U_C} = 1$

[10] An example of a right-nested counterfactual will be discussed in Section 5. Note that left-nested counterfactuals are often excluded from counterfactual analysis, c.f. Briggs (2012); this issue is also discussed in Vandenburgh (2021).

are the only two interventions setting the antecedent true in $I_u(A)$. Then, for $X = 1 > D = 1$ to be true, we need to verify that $D = 1$ is true in the two intervened worlds where we intervene on $(U_C, U_X, U_Y) = (0, 0, 0)$ by setting $U_X = 1$ and $U_C = 1$. These intervened worlds are $(U_C, U_X, U_Y) = (0, 1, 0)$ and $(U_C, U_X, U_Y) = (1, 0, 0)$: in both cases, $X = 1$, so since $D = X \vee Y$, $D = 1$, guaranteeing that the counterfactual is true, as expected.

The exogenous intervention model can also account for backtracking counterfactuals, unlike Pearl's interventions through structural equations. Consider the backtracking counterfactual "If the prisoner had died, either X or Y would have shot." The exogenous interventions setting $D = 1$ are $U_C = 1$, $U_X = 1$, and $U_Y = 1$. All of these interventions ensure that $X = 1 \vee Y = 1$ is true: $U_X = 1$ sets $X = 1$, $U_Y = 1$ sets $Y = 1$, and $U_C = 1$ sets both $X = 1$ and $Y = 1$. Thus, $X = 1 \vee Y = 1$ is true in all intervened worlds in the selection function, so the backtracking counterfactual "If the prisoner had died, either X or Y would have shot" is true. This verdict is intuitive: in the model, the only way the prisoner could have died is if either X or Y shot the prisoner.

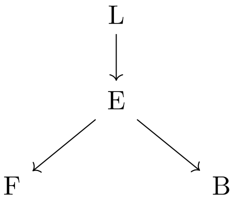
4 Restricting the Selection Function: Hiddleston

The exogenous intervention model uses the causal notion of intervention to define a selection function for counterfactual semantics: $A > C$ is true if C is true in all worlds where one has set A true through an A-intervention. While the structure mirrors that of the similarity theory of counterfactuals, the domain for counterfactual evaluation is determined by causal considerations rather than intuitions about similarity. Imagine that a student receives a score of 95/100 on an essay, and consider the counterfactual "If the student had scored lower, she would have scored 94." Intuitively, this counterfactual is false, but a selection function built on similarity might very well judge it true: the world where the student scores 94 is closer to the actual world than worlds where the student scores 93 or below, making it the unique closest world satisfying the antecedent. However, this is not the case for the selection function in Section 3: the instructor could have intervened to set the score to any value, like 93 or 92, or even 65. These are all causal interventions, and the model in Section 3 invokes no restriction on the selection function ruling out "more distant" or "less likely" interventions.

While the exogenous intervention model from Section 3 includes all possible exogenous interventions for the selection function, regardless of how significantly the interventions change the actual world, it is straightforward to modify the model to consider only a restricted set of interventions. If one has a distance metric or similarity ordering on the set of causal worlds $\mathcal{U}$, then one can restrict the set of intervened worlds in the selection function from Section 3 to only the closest or most similar worlds. While this procedure applies to any distance measure or similarity ordering, I focus on causal restrictions to the set of relevant interventions rather than invoking intuitions about similarity, which can restrict to too few worlds, as above, or lead to problematic judgments like in Fine's case.

One such causal restriction comes from Eric Hiddleston, who argues that counterfactuals should consider only interventions that occur as late as possible in the causal process. This requirement will also appear in the exogenous interpretation of Pearl's approach to interventions in the next section. Before discussing how we can implement this restriction within the selection function from Section 3, it is worth discussing how the formal set-up of the exogenous intervention model differs from that of Hiddleston, which will highlight some advantages which come from utilizing exogenous interventions.

Hiddleston's theory follows the set-up of Section 2 with two fundamental differences: he considers all variables as endogenous and he allows for indeterministic structural equations such as $\Pr(\mathrm{Y} = y|\mathrm{X} = x) = p$, where $p \in (0, 1)$. To see how Hiddleston's framework works, consider his 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 $\Pr(\mathrm{E} = 1|\mathrm{L} = 1) = 0.95$, $\Pr(\mathrm{E} = 1|\mathrm{L} = 0) = 0$, $\mathrm{F} = \mathrm{E}$, and $\mathrm{B} = \mathrm{E}$ with causal graph:



Hiddleston evaluates a counterfactual $\mathrm{A} > \mathrm{C}$ at u by considering whether C is true in all models which are "minimal breaks" from the model in u, where a break is a change in the values of endogenous variables consistent with the causal laws. For example, in evaluating the counterfactual "If the flash hadn't occurred, the cannon would not have exploded," causal breaks include the break where the cannon is not lit ($\mathrm{L} = 0$) as well as the break where the explosion does not happen ($\mathrm{E} = 0$). The former is consistent with the causal laws because L has no parents, and so can be set without constraint, and the latter is consistent with the causal laws because the law relating E and L is indeterministic: it is perfectly consistent with the structural equations that the cannon is lit ($\mathrm{L} = 1$) but does not explode ($\mathrm{E} = 0$).

Using indeterministic structural equations, however, can lead to too many causal breaks to account for counterfactual judgments, especially for forward counterfactuals. Consider the forward counterfactual "If the cannon had been lit, then an explosion would have happened." Despite the fact that the explosion does not result from lighting the cannon 100% of the time, it is plausible that this counterfactual is true: in all normal or typical situations, lighting a cannon produces an explosion. For Hiddleston, however, this counterfactual must be false, since there is always a possibility that the explosion does not result from lighting the cannon. This problem arises for other counterfactuals as well: we often judge counterfactuals like “If I had gone to the bakery, I would have bought bread” true, even though there is a small chance that the bakery is out of bread.

Exogenous interventions can capture the causal breaks Hiddleston is interested in, including those from indeterministic structural equations, without introducing too much indeterminacy into forward counterfactuals. We can accomplish this by translating Hiddleston’s examples into the terminology of Section 2 in two steps. First, we can add an exogenous variable determining any variable without parents: if $\mathrm{PV}_i = \emptyset$, we can add an exogenous variable U_i with the same variable values as V_i such that the structural equation for V_i is $\mathrm{V}_i = \mathrm{U}_i$. Second, we can account for indeterministic structural equations by introducing exogenous error variables, which capture ways in which causal outcomes can deviate from the outcomes expected based on the parent variable values. Common error terms for a binary variable include “unspecified inhibiting abnormalities,” or things which prevent the parents from activating the variable, and “unspecified triggering abnormalities,” or things which trigger the variable independent of the parents.[11] For example, we can account for the indeterministic relationship between L and E by adding an error variable U'_{E} representing inhibiting abnormalities for the explosion, yielding a deterministic structural equation $\mathrm{E} = \mathrm{L} \wedge \neg \mathrm{U}'_{\mathrm{E}}$ which says that E is activated when L is activated and L is not inhibited by U'_{E}. Here, 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 $\Pr(\mathrm{U}'_{\mathrm{E}} = 1) = .05$. Thus, putting the ceremonial cannon case into the terminology from Section 2, there are two exogenous variables, U_{L} and U'_{E}, with endogenous variables and causal graph as above and with structural equations $\mathrm{L} = \mathrm{U}_{\mathrm{L}}$, $\mathrm{E} = \mathrm{L} \wedge \neg \mathrm{U}'_{\mathrm{E}}$, $\mathrm{F} = \mathrm{E}$, and $\mathrm{B} = \mathrm{E}$.

Using exogenous interventions in the translated causal structure avoids the problem raised for forward counterfactuals with indeterministic structural equations. Recall the counterfactual “If the cannon had been lit, then an explosion would have happened.” With indeterministic structural equations, this is always false: there is always a 5% chance that an explosion would not have happened. But with exogenous interventions, the counterfactual is true in worlds where the error term U'_{E} is inactive, which is the normal situation occurring 95% of the time, and the counterfactual is false in the outlier worlds where the error term is activated. Thus, using exogenous interventions means that counterfactuals can be true even with a small probability of something going wrong, a fact which is important for a theory of counterfactuals to capture.[12]

[11] For Pearl’s discussion of error variables in Boolean models, see Pearl (2009, pp. 29).

[12] Note that Edgington (1995, 2008) argues that any universally quantified theory of counterfactuals is unable to account for probabilistic exceptions, arguing instead that the acceptability of a counterfactual $A > C$ should be based on the probability of the consequent C conditional on the antecedent A. Part of the motivation for this is that universally quantified theories predict that a single improbable exception can completely falsify a counterfactual, while on a probabilistic account, such an exception merely lowers the probability appropriately. However, I think that the universally quantified theory is necessary to explain some intuitions about counterfactuals, such as the unacceptability of certain counterfactuals with single improbable exceptions like “If I had bought a lottery ticket, it would have lost.” Furthermore, I think there are promising avenues for reconciling the universally quantified theory with intuitions

Now that we have discussed the formal differences between Hiddleston's theory of causal breaks and exogenous interventions, we can introduce Hiddleston's additional minimality constraint. For Hiddleston, the only causal breaks we should consider to set a counterfactual antecedent true are those making changes to variables which are "as minor and as late as is lawfully possible" (Hiddleston, 2005, pp. 643). The condition that changes are as minor as possible is enforced by the notion of an exogenous intervention: exogenous interventions are limited to variable changes which are necessary to set the antecedent true, so no exogenous intervention could set the antecedent true with only a subset of its variable changes. However, we have made no requirement that changes be as late as possible in the causal process. Such a requirement could even be seen as a natural generalization of the requirement that an intervention not change any unnecessary exogenous variables: an intervention which is as late as possible does not change any unnecessary endogenous variables.

Consider the following counterfactual: "If the flash hadn't occurred, the cannon would still have been lit," and assume the world is such that the cannon is lit and the flash occurred, represented as $(\mathrm{U_L}, \mathrm{U'_E}) = (1, 0)$. Here, the two exogenous interventions which could set the antecedent true are $\mathrm{U_L} = 0$, where the cannon is not lit, and $\mathrm{U'_E} = 1$, where an abnormality intervenes so that, even though the cannon is lit, it does not explode and, consequently, there is no flash. In the theory from Section 3, both of these are equally good exogenous interventions fixing the antecedent, so the counterfactual is false, as the intervention $\mathrm{U_L} = 0$ prevents the cannon from being lit. Hiddleston argues, however, that the counterfactual should be true: the intervention setting $\mathrm{U_L} = 0$ is earlier in the causal process and unnecessarily changes the value of the endogenous variable L, so the only relevant intervention is $\mathrm{U'_E} = 1$, on which the cannon is still lit. I disagree with Hiddleston's judgment in this case, as well as his reasoning for restricting to interventions which are as late as possible: the most likely explanation for why the flash would not have occurred is that the cannon was not lit, so it seems misguided to remove this possibility simply because it changes more facts in the actual world. However, Hiddleston's constraint can easily be incorporated in the exogenous intervention model if desired. We can define an ordering $\leq_{\mathrm{H}}$ on interventions $\mathrm{I}_u(\mathrm{A})$ such that $i_{\mathrm{S}} \leq_{\mathrm{H}} i'_{\mathrm{S'}}$ if the set of endogenous variables i_{S} changes in u is a subset of the set of endogenous variables $i'_{\mathrm{S'}}$ changes in u. Enforcing Hiddleston's constraint, a counterfactual $\mathrm{A} > \mathrm{C}$ is true in u if C is true in all worlds reached from u by a $\leq_{\mathrm{H}}$-minimal A-intervention. This translation of Hiddleston's theory into the exogenous intervention framework highlights one way to restrict the framework from Section 3 with a more substantial requirement of minimality or similarity.

about the probabilities of conditionals and counterfactuals; see Egré and Cozic (2011) and Vandenburgh (2020).

5 Comparison to Endogenous Interventions: Pearl and Briggs

I have argued that the ability to handle backtracking counterfactuals is a virtue of the exogenous intervention model, and this virtue is shared by other models where counterfactual intervention involves changing the values of variables, like Hiddleston's model. This, however, diverges from Pearl's theory of counterfactuals, where interventions change structural equations rather than variable values and backtracking reasoning is excluded. For Pearl, the exclusion of backtracking reasoning is intentional: permitting backtracking reasoning in a theory of interventions 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 determined by the 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 a change to 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. We can think of this exogenous variable 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.[13] 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.[14] The fact that economists estimate this exogenous effect on interest rates and incorporate an exogenous variable representing it in their models provides evidence in favor of adding this variable to the model, allowing exogenous interventions to avoid the backtracking problem.

This argument extends to other cases where the backtracking problem may arise: whenever one can intervene on a variable, one can add an exogenous variable to capture this possibility. This is related to the procedure of adding intervention variables to causal models in Meek and Glymour (1994), which allows one to intervene by conditioning on variable values rather than by changing structural equations. One difference between exogenous intervention and Pearl's theory, however, is that Pearl allows for direct interventions on any variable in a model, regardless of whether there is a real or salient possibility of an exogenous influence on that variable (Pearl, 2009, pp. 361). For example, if one has an unplugged microwave, Pearl's theory allows one to intervene to turn the microwave on without requiring that the microwave be plugged in (which requires backtracking reasoning), even though there is no real or salient possibility of this occurring. While I think such cases motivate the need to incorporate backtracking in counterfactual reasoning, as argued in Section 2, this is not required by the exogenous intervention model itself: the notion of an exogenous intervention is general enough to represent Pearl's theory of "intervention without manipulation." Thus, exogenous interventions are more general than endogenous interventions on structural equations: exogenous interventions can imitate interventions on structural equations, but there is no way to modify interventions on structural equations to accommodate backtracking reasoning.

To capture Pearl's theory of interventions with exogenous interventions, we would need to, first, add an exogenous influence to every endogenous variable in the model and, second, restrict the set of interventions to those which do no change any parent variables. For the first step, we can create a new causal model where, for any endogenous variable which does not include the possibility of an exogenous shock, we simply add an exogenous variable $\mathrm{U}_i = \mathrm{V}_i \cup \{\mathrm{OFF}\}$, where V_i is determined according to its original structural equation when $\mathrm{U}_i = \mathrm{OFF}$ and $\mathrm{V}_i = \mathrm{U}_i$ otherwise. For the second step, we can restrict to the $\leq_{\mathrm{H}}$-minimal interventions in the new causal model. Thus, when evaluating a counterfactual $\mathrm{A} > \mathrm{C}$ in a causal model $\mathcal{M}$, we can approximate Pearl's predictions by looking at the $\leq_{\mathrm{H}}$-minimal interventions in the expanded model $\mathcal{M}'$ with additional

[13] 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).

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

exogenous variables. Since $\mathcal{M}'$ includes exogenous influences for all variables, there is an intervention in $\mathrm{I}_u(\mathrm{A})$ influencing the variables in A directly, and any intervention changing variables upstream from A is an "earlier" intervention eliminated by the $\leq_{\mathrm{H}}$-minimality condition, so the only counterfactually relevant interventions are those directly influencing variables from A, as in Pearl's theory.

Even following these steps to approximate the predictions of Pearl's theory, however, some differences arise for logically complex counterfactuals. One difference involves counterfactuals where the antecedent A is actual. On Pearl's theory of interventions, even though A is true in the actual world, an intervention setting A true still requires us to change the structural equations for A. This leads to violations of modus ponens when Pearl's theory is extended to logically complex counterfactuals, as shown in Briggs (2012). On the exogenous intervention model, however, the idea of intervening to bring about a fact which is already true makes little sense: the only exogenous A-intervention is the empty intervention. This property, corresponding to the logical principle of strong centering, guarantees that modus ponens is valid.

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 $\mathrm{U_C}, \mathrm{U_X}$, and $\mathrm{U_Y}$ is sufficiently rich. Consider the nested counterfactual "If X had shot, then if the court had not ordered it, the prisoner would have died," represented $\mathrm{X} > (\neg\mathrm{C} > \mathrm{D})$.[15] Assume the world is $(\mathrm{U_C}, \mathrm{U_X}, \mathrm{U_Y}) = (1, 0, 0)$, where the court orders execution, the signal is sent, X and Y both shoot, and the prisoner dies. The nested counterfactual $\mathrm{X} > (\neg\mathrm{C} > \mathrm{D})$ is false in this world on the exogenous intervention model: since X is true in the actual world, the only relevant intervened X-world is the actual world, and in this world, the consequent is false, as the only intervention $\mathrm{U_C} = 0$ setting $\neg$C true prevents the prisoner from dying. This verdict is intuitive: the second intervention ensuring that the court did not order execution overrides the initial supposition that X had shot on the basis of the court's order, so the relevant intervened world is one where the court does not order execution and X does not shoot, and, consequently, the prisoner does not die. Furthermore, this counterfactual must be false for modus ponens to be true: if $\mathrm{X} > (\neg\mathrm{C} > \mathrm{D})$ were true, then since X is true in the actual world, modus ponens entails that $\neg\mathrm{C} > \mathrm{D}$ is true, but $\neg\mathrm{C} > \mathrm{D}$ is false because the intervention setting $\mathrm{C} = 0$ also sets $\mathrm{D} = 0$.[16] In Briggs' extension of Pearl's theory, however,

[15] Adapted from Briggs (2012, pp. 150). For convenience, we use the variable names V as shorthand for the proposition that the variable is activated, $\mathrm{V} = 1$, and $\neg$V for the negation, that $\mathrm{V} = 0$.

[16] Note that some authors, like McGee (1985), argue that right-nested conditionals are genuine counterexamples to modus ponens. These counterexamples are, however, framed as indicative conditionals rather than counterfactuals, and counterfactual versions of these cases may be less compelling. For example, McGee's argument against modus ponens relies on his claim that the following indicative conditional about the 1980 presidential election is true: "If a Republican wins, then if Reagan does not win, Anderson will win," as Anderson was the leading Republican behind Reagan. However, the counterfactual "If a Republican had won, then if Reagan hadn't won, Anderson would have won" strikes me as false: the second

the nested counterfactual $\mathrm{X} > (\neg\mathrm{C} > \mathrm{D})$ is true, violating modus ponens. This is because intervening to set X true changes the structural equation for X from $\mathrm{X} = \mathrm{S} \vee \mathrm{U_X}$ to $\mathrm{X} = 1$: when we subsequently evaluate $\neg\mathrm{C} > \mathrm{D}$ by replacing $\mathrm{C} = \mathrm{U_C}$ with $\mathrm{C} = 0$, the fact that the structural equation for X is still $\mathrm{X} = 1$ ensures that $\mathrm{D} = 1$, so $\mathrm{X} > (\neg\mathrm{C} > \mathrm{D})$ is true.

Another difference arises for disjunctive antecedents. Imagine, following an example from Ciardelli et al. (2018), that a light L is controlled by two switches, $\mathrm{S_1}$ and $\mathrm{S_2}$, where the light is on when both switches are in the same position (i.e., both are up or both are down): $\mathrm{L} = (\mathrm{S_1} \wedge \mathrm{S_2}) \vee (\neg\mathrm{S_1} \wedge \neg\mathrm{S_2})$. Imagine both switches are up and the light is on, $(\mathrm{S_1}, \mathrm{S_2}) = (1, 1)$, and consider the counterfactual "If $\mathrm{S_1}$ or $\mathrm{S_2}$ had been down, the light would have been off." Intuitively, and based on the experimental responses presented by Ciardelli et al. (2018), this counterfactual is true. This is explained by the exogenous intervention model, as the exogenous interventions setting the antecedent true are $\mathrm{S_1} = 0$ and $\mathrm{S_2} = 0$, and in both intervened worlds, one switch is up and one is down, so the light is always off. The combined intervention $(\mathrm{S_1}, \mathrm{S_2}) = (0, 0)$ is not an exogenous intervention setting the antecedent true because it makes irrelevant changes: either changing $\mathrm{S_1}$ or changing $\mathrm{S_2}$ is sufficient to set the antecedent true, so there is no need to consider the combined intervention.[17] On Pearl's theory, however, there is no principle excluding the combined intervention for disjunctive antecedents. While Pearl's original theory does not incorporate disjunctive antecedents, Briggs' extension of Pearl's theory includes the combined intervention, as does the theory in Santorio (2019) inspired by Pearl's approach. Thus, while the correct prediction follows directly from the exogenous intervention model, it need not (and often does not) arise in models based on Pearl's theory of endogenous interventions.[18]

6 Conclusion

In this paper, I argued for the use of exogenous interventions to capture the semantics of counterfactual sentences. On this approach, a counterfactual $\mathrm{A} > \mathrm{C}$ 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 incorporates both forward and backtracking counterfactuals, applies to logically complex antecedents, and satisfies the axioms of a familiar counterfactual logic, Pollock's

intervention cancels the first intervention, so the relevant intervened world is one where Reagan hadn't won and a Republican need not have won, and in this world, Carter could have (and likely would have) won.

[17] This corresponds to the logical principle CS5′ satisfied by the exogenous intervention model, as shown in the Appendix.

[18] Note that Ciardelli et al. (2018) argue that classical counterfactual theories built on selection functions, like the exogenous intervention model, are also inadequate for explaining all logically complex counterfactuals. For example, the counterfactual "If $\mathrm{S_1}$ and $\mathrm{S_2}$ had not both been up, the light would have been off" is classically equivalent to the above counterfactual, but is often judged false. However, some recent work suggests this can be addressed with a more nuanced understanding of negation (Schulz, 2018; Romoli et al., 2022).

SS. This approach can be extended by considering additional restrictions on the selection function, as illustrated in the reformulation of Hiddleston's theory in Section 4, and can capture many of the intuitions of Pearl's approach to counterfactuals, provided the model includes sufficiently many exogenous variables. Exogenous intervention therefore offers a flexible approach to counterfactual reasoning which can combine interventionist intuitions with the logical advantages of the similarity theory of counterfactuals.

References

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

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

Christiano, L. J., Eichenbaum, M., & 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., & Champollion, L. (2018). Two switches in the theory of counterfactuals. *Linguistics and Philosophy*, 41(6), 577–621.

Edgington, D. (1995). On conditionals. *Mind*, 104(414), 235–329.

Edgington, D. (2008). I—counterfactuals. In *Proceedings of the Aristotelian Society*, volume 108 (pp. 1–21). Oxford: Blackwell Publishing Ltd.

Egré, P. & Cozic, M. (2011). If-clauses and probability operators. *Topoi*, 30(1), 17–29.

Fine, K. (1975). Critical notice. *Mind*, 84(335), 451–458.

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

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

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

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35 (pp. 2386–2391).

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

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

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

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

Halpern, J. Y. & Hitchcock, C. (2010). Actual causation and the art of modeling. In *Heuristics, Probability, and Causality: A Tribute to Judea Pearl* (pp. 383–406). New York, NY: College Publications.

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* (pp. 220–229). New York, NY: Springer.

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

LeRoy, S. F. (2020). *Causal Inference in Economic Models*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, (pp. 455–476).

Lewis, D. (2013). *Counterfactuals*. Hoboken, NJ: John Wiley and Sons.

Lucas, C. & Kemp, C. (2012). A unified theory of counterfactual reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34 (pp. 707–712).

Lucas, C. & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, 122(4), 700–734.

McGee, V. (1985). A counterexample to modus ponens. *The Journal of Philosophy*, 82(9), 462–471.

Meek, C. & Glymour, C. (1994). Conditioning and intervening. *The British Journal for the Philosophy of Science*, 45(4), 1001–1021.

Pearl, J. (2009). *Causality.* Cambridge: 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.

Romoli, J., Santorio, P., & Wittenberg, E. (2022). Alternatives in counterfactuals: What is right and what is not. *Journal of Semantics*, 39(2), 213–260.

Santorio, P. (2019). Interventions in premise semantics. *Philosophers' Imprint*, 19(1), 1–27.

Schulz, K. (2018). The similarity approach strikes back: Negation in counterfactuals. In *Proceedings of Sinn und Bedeutung*, volume 22 (pp. 343–360).

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

Stalnaker, R. (1968). A theory of conditionals. In *Ifs* (pp. 41–55). New York, NY: Springer.

Vandenburgh, J. (2020). Triviality results, conditional probability, and restrictor conditionals. Available at https://philpapers.org/rec/VANTRC-4.

Vandenburgh, J. (2021). Conditional learning through causal models. *Synthese*, 199(1), 2415–2437.

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

1 Appendix: Logic of Exogenous Intervention Models

The exogenous intervention model defines counterfactual truth conditions in terms of a selection function: a counterfactual $\mathrm{A} > \mathrm{C}$ is true in a world u if C is true in all worlds in $f(\mathrm{A}, u)$ reached from u through an exogenous A-intervention. This model satisfies the axioms of Pollock's (1981) counterfactual logic SS, corresponding to six axioms for the selection function:

CS1: if $w \in f(\mathrm{A}, u)$, then $w \in [\mathrm{A}]$
CS2: if $u \in [\mathrm{A}]$, then $f(\mathrm{A}, u) = \{u\}$
CS3: if $f(\mathrm{A}, u) = \emptyset$, then $f(\mathrm{B}, u) \cap [\mathrm{A}] = \emptyset$
CS4: if $f(\mathrm{A}, u) \subseteq [\mathrm{B}]$ and $f(\mathrm{B}, u) \subseteq [\mathrm{A}]$, then $f(\mathrm{A}, u) = f(\mathrm{B}, u)$
CS5′: $f(\mathrm{A} \vee \mathrm{B}, u) \subseteq f(\mathrm{A}, u) \cup f(\mathrm{B}, u)$
CS6: $u \in [\mathrm{A} > \mathrm{C}]$ iff $f(\mathrm{A}, u) \subseteq [\mathrm{C}]$

We verify that the selection function for the exogenous intervention model in Section 3 satisfies these six axioms for Pollock's logic SS below:

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

Proof. Suppose $w \in f(\mathrm{A}, u)$, so $w = u|i_\mathrm{S}$ for some $i_\mathrm{S} \in \mathrm{I}_u(\mathrm{A})$. Since $i_\mathrm{S} \in \mathrm{R}_u(\mathrm{A})$, $u|i_\mathrm{S} \in [\mathrm{A}]$ by the definition of $\mathrm{R}_u(\mathrm{A})$, so $w \in [\mathrm{A}]$. □

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

Proof. If $u \in [\mathrm{A}]$, then the empty intervention i_0, which changes no exogenous variables, is in $\mathrm{R}_u(\mathrm{A})$ since $u|i_0 = u \in \mathrm{R}_u(\mathrm{A})$. Since $i_0 \leq r_\mathrm{S}$ for every other possible intervention $r_\mathrm{S} \in \mathrm{R}_u(\mathrm{A})$, i_0 is the unique $\leq$-minimal element in $\mathrm{R}_u(\mathrm{A})$ and the only element in $\mathrm{I}_u(\mathrm{A})$. Since $f(\mathrm{A}, u) = \{u|i_\mathrm{S} : i_\mathrm{S} \in \mathrm{I}_u(\mathrm{A})\}$, $f(\mathrm{A}, u) = \{u|i_0\} = \{u\}$. □

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

Proof. If $f(\mathrm{A}, u) = \emptyset$, then $\mathrm{I}_u(\mathrm{A}) = \emptyset$, so $\mathrm{R}_u(\mathrm{A}) = \emptyset$. Since $[\mathrm{A}] \subseteq \mathrm{R}_u(\mathrm{A})$, $[\mathrm{A}] = \emptyset$, so $f(\mathrm{B}, u) \cap [\mathrm{A}] = \emptyset$. □

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

Proof. Suppose $f(\mathrm{A}, u) \subseteq [\mathrm{B}]$ and $f(\mathrm{B}, u) \subseteq [\mathrm{A}]$. To show that $f(\mathrm{A}, u) \subseteq f(\mathrm{B}, u)$, we must show that, for all $i_\mathrm{S} \in \mathrm{I}_u(\mathrm{A})$, there is some $j_{\mathrm{S}^*} \in \mathrm{I}_u(\mathrm{B})$ such that $u|i_\mathrm{S} = u|j_{\mathrm{S}^*}$. Suppose $i_\mathrm{S} \in \mathrm{I}_u(\mathrm{A})$. Since $f(\mathrm{A}, u) \subseteq [\mathrm{B}]$, $u|i_\mathrm{S} \in [\mathrm{B}]$, so $i_\mathrm{S} \in \mathrm{R}_u(\mathrm{B})$. Then there is a $j_{\mathrm{S}^*} \in \mathrm{I}_u(\mathrm{B})$ such that i_S extends j_{S^*}. But since $j_{\mathrm{S}^*} \in \mathrm{I}_u(\mathrm{B})$ and $f(\mathrm{B}, u) \subseteq [\mathrm{A}]$, $u|j_{\mathrm{S}^*} \in [\mathrm{A}]$, so $j_{\mathrm{S}^*} \in \mathrm{R}_u(\mathrm{A})$. This means there is an $i'_{\mathrm{S}'} \in \mathrm{I}_u(\mathrm{A})$ such that j_{S^*} extends $i'_{\mathrm{S}'}$. But since i_S and $i'_{\mathrm{S}'}$ are both $\leq$-minimal elements and $i'_{\mathrm{S}'} \leq j_{\mathrm{S}^*} \leq i_\mathrm{S}$, $i_\mathrm{S} = i'_{\mathrm{S}'} = j_{\mathrm{S}^*}$, so $u|i_\mathrm{S} = u|j_{\mathrm{S}^*}$. Since we have shown $\forall i_\mathrm{S} \in \mathrm{I}_u(\mathrm{A}), \exists j_{\mathrm{S}^*} \in \mathrm{I}_u(\mathrm{B})$ such that $u|i_\mathrm{S} = u|j_{\mathrm{S}^*}$, we have shown that $f(\mathrm{A}, u) \subseteq f(\mathrm{B}, u)$. The proof that $f(\mathrm{B}, u) \subseteq f(\mathrm{A}, u)$ is parallel, showing that $f(\mathrm{A}, u) = f(\mathrm{B}, u)$. □

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

Proof. Suppose $u|i_{\mathrm{S}} \in f(\mathrm{A} \vee \mathrm{B}, u)$, where $i_{\mathrm{S}} \in \mathrm{I}_u(\mathrm{A} \vee \mathrm{B})$. Since $u|i_{\mathrm{S}} \in [\mathrm{A} \vee \mathrm{B}]$ by CS1, $u|i_{\mathrm{S}} \in [\mathrm{A}]$ or $u|i_{\mathrm{S}} \in [\mathrm{B}]$. Suppose $u|i_{\mathrm{S}} \in [\mathrm{A}]$. Then $i_{\mathrm{S}} \in \mathrm{R}_u(\mathrm{A})$, so there is some $j_{\mathrm{S}^*} \in \mathrm{I}_u(\mathrm{A})$ such that i_{S} extends j_{S^*}. Since $j_{\mathrm{S}^*} \in \mathrm{I}_u(\mathrm{A})$, $u|j_{\mathrm{S}^*} \in [\mathrm{A}] \subseteq [\mathrm{A} \vee \mathrm{B}]$, so $j_{\mathrm{S}^*} \in \mathrm{R}_u(\mathrm{A} \vee \mathrm{B})$. This means there is some $i'_{\mathrm{S}'} \in \mathrm{I}_u(\mathrm{A} \vee \mathrm{B})$ such that j_{S^*} extends $i'_{\mathrm{S}'}$. But since $i'_{\mathrm{S}'} \leq j_{\mathrm{S}^*} \leq i_{\mathrm{S}}$ and i_{S} and $i'_{\mathrm{S}'}$ are both $\leq$-minimal, $i_{\mathrm{S}} = i'_{\mathrm{S}'} = j_{\mathrm{S}^*}$, so $\exists j_{\mathrm{S}^*} \in \mathrm{I}_u(\mathrm{A})$ such that $u|i_{\mathrm{S}} = u|j_{\mathrm{S}^*}$, so $u|i_{\mathrm{S}} \in f(\mathrm{A}, u) \cup f(\mathrm{B}, u)$. If $u|i_{\mathrm{S}} \in [\mathrm{B}]$, a parallel proof shows that $u|i_{\mathrm{S}} \in f(\mathrm{B}, u) \subseteq f(\mathrm{A}, u) \cup f(\mathrm{B}, u)$. Therefore, $f(\mathrm{A} \vee \mathrm{B}, u) \subseteq f(\mathrm{A}, u) \cup f(\mathrm{B}, u)$. □

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

Proof. Follows immediately from the definition of $[\mathrm{A} > \mathrm{C}]$ in Section 3. □

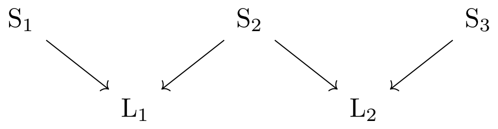
Pollock's logic SS is slightly weaker than Lewis's (2013) logic VC, which replaces CS5′ with CS5:

CS5: if $f(\mathrm{A}, u) \cap [\mathrm{B}] \neq \emptyset$, then $f(\mathrm{A} \wedge \mathrm{B}, u) \subseteq f(\mathrm{A}, u)$.

The exogenous intervention model does not validate the stronger axioms of VC, as it admits counterexamples to CS5 and the corresponding logical principle:

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

We can see this with a standard counterexample 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 $\mathrm{L}_1 = \mathrm{S}_1 \vee \mathrm{S}_2$ and $\mathrm{L}_2 = \mathrm{S}_2 \vee \mathrm{S}_3$. The causal diagram for this model is as follows:



Suppose all three switches are off ($\mathrm{S}_i = 0$) and, consequently, both lights are off ($\mathrm{L}_i = 0$). The counterfactual "If L_2 had been on, S_1 would have been off" is true since both interventions which set $\mathrm{L}_2 = 1$, $\mathrm{S}_2 = 1$ and $\mathrm{S}_3 = 1$, leave S_1 fixed at 0. Additionally, it is not the case that "If L_2 had been on, L_1 would have been on," since setting $\mathrm{S}_3 = 1$ is an intervention which fixes $\mathrm{L}_2 = 1$ without setting $\mathrm{L}_1 = 1$. However, it is not the case that "If L_1 and L_2 had been on, S_1 would have been off." Here, the two interventions setting the antecedent true are $\mathrm{S}_2 = 1$ and $(\mathrm{S}_1, \mathrm{S}_3) = (1, 1)$, and the latter intervention leaves S_1 on, so the counterfactual is false. This counterexample to the logical principle corresponding to CS5 shows that the exogenous intervention model does not validate Lewis's logic VC without additional restrictions on the selection function.