Causality Without Causal Models

Joseph Y. Halpern
Cornell University
halpern@cs.cornell.edu

Rafael Pass
Cornell Tech, Technion, TAU
rafael@cs.cornell.edu

Perhaps the most prominent current definition of (actual) causality is due to Halpern and Pearl. It is defined using causal models (also known as *structural equations models*). We abstract the definition, extracting its key features, so that it can be applied to any other model where counterfactuals are defined. By abstracting the definition, we gain a number of benefits. Not only can we apply the definition in a wider range of models, including ones that allow, for example, backtracking, but we can apply the definition to determine if A is a cause of B even if A and B are formulas involving disjunctions, negations, beliefs, and nested counterfactuals (none of which can be handled by the Halpern-Pearl definition). Moreover, we can extend the ideas to getting an abstract definition of explanation that can be applied beyond causal models. Finally, we gain a deeper understanding of features of the definition even in causal models.

1 Introduction

Perhaps the most prominent current definition of (actual) causality is due to Halpern and Pearl [7, 8] (HP from now on). It is defined using causal models (also known as *structural equations models*). The fact that the definition is so tied to causal models has benefits—for example, the fact that causal models can be represented graphically makes them relatively easy to work with—but also leads to a number of limitations; specifically, the language in which we can express counterfactuals is limited (so we cannot express causal statements that involve more complex counterfactuals), and we cannot deal with *backtracking*. (See below for further discussion.) The goal of this paper is to abstract the definition of causality in causal models, extracting its key features, so that it can be applied to any other model where counterfactuals are defined.

Doing so lets us get around these limitations. For one thing, it lets us consider more general causal statements. The HP definition of "A is a cause of B" in causal models can be applied only when A is a conjunction of primitive events $X = x$, where X is a variable and x is a value in the range of X, and B is a Boolean combination of primitive events. It cannot be applied if A is a disjunction, or if either A or B involves modalities like belief or nested counterfactuals. But such statements are important in practice! For example, Sartorio [19] gives examples where disjunctive causes play a significant role; Soloviev and Halpern [20] give examples where nested counterfactuals are critical for expressing security properties. And we clearly want to be able to reason about (and intervene on) agent's beliefs. For example, we might want to say "If were to intervene on Alice's beliefs so that she believed that the vaccine was effective, then she would take it".

Causal models have been extended to allow counterfactuals that involve disjunctions and nested counterfactuals [2], and reasoning about both knowledge/belief and causality [1] (although not interventions on beliefs). These extensions involving disjunction and nested counterfactuals are not at all straightforward, and lead to some arguably unreasonable axioms. For example, the axiom given by Briggs that characterizes interventions involving disjunctions is $[A \vee A']B \Leftrightarrow ([A]B \wedge [A']B \wedge [A \wedge A']B)$. It is not clear why this is appropriate. Briggs does not try to justify it and, indeed, admits that the axiom validates "the controversial

Submitted to:
TARK 2025

rule *Simplification of Disjunctive Antecedents* ... [which] most authors reject". Dealing with interventions on beliefs also seems quite nontrivial. All this means that a definition of causality involving these constructs may well be controversial. The advantage of the approach that we are suggesting here is that, as long as we start with a framework that has a definition of counterfactuals that a user is comfortable with and supports these constructs, we can essentially import the HP definition to that framework. In particular, we get a definition of causality in the counterfactual structures considered by Stalnaker [21] and Lewis [14] that are standard in the philosophy literature. Moreover, we can show that this definition generalizes that in causal models: as shown by Halpern [6], causal models correspond to a subset of counterfactuals; we show here (see Theorem 4.5) that, in a causal model M, the original definition of causality agrees with the new definition in the counterfactual structures M' corresponding to M.

As we said above, the second key limitation of defining causality in causal models is that we cannot deal with backtracking. As has often been observed, when considering the effect of an intervention such as $X = x$ in a causal model, no variables "upstream" of X in the causal graph are affected; only descendants of X can be affected. This is referred to as "no backtracking".

The argument for no backtracking is not hard to explain. If we intervene to set, say $X = 2$, then, intuitively, all that should change are the values of variables that are descendants of X, since these are the only ones that can be affected by the intervention. But, as argued by von Kügelen et al. [13], there are cases where we want to ascribe causality but do not want to consider counterfactuals that result from interventions. Intuitively, these are cases where the causal laws, not the background conditions, are shared between the actual and counterfactual worlds. Consequently, the "upstream" variables must be allowed to differ to accommodate possibly contradictory facts. As discussed by von Kügelen et al. (who also give references to work in the cognitive psychology literature on the topic), people certainly consider backtracking counterfactuals when evaluating causality. This makes it useful to have a definition of causality that allows them. Causal models have been extended so as to deal with backtracking [13, 16], but the definitions of causality used are quite different from the HP definition. Our approach lets us import the HP definition "for free", so to speak, as long as we start with a definition of counterfactual that allows backtracking (again, as do the standard definitions of Stalnaker [21] and Lewis [14]).

This approach can be further extended to give an abstract definition of *explanation*. There are many definitions of explanations in the literature. The modern literature goes back to the work of Hempel [10] and Salmon [17], but this work is well known to have problems because it does not take causality into account. Here we focus on causal explanations, again taking as our basis the work of Halpern and Pearl [7, 9]. The definition is based on the defintion of causality (which is what will allow us to apply the approach of this paper), but it also takes into account the well-known observation that what counts as an explanation is relative to what an agent knows [3, 18]. As Gärdenfors [3] observes, an agent seeking an explanation of why Mr. Johansson has been taken ill with lung cancer will not consider the fact that he worked for years in asbestos manufacturing an explanation if he already knew this fact. It is relatively straightforward to take an agent's knowledge into account in our abstract approach. Doing so gives us the tools we need to handle explanation as well.

Our Definition in a Nutshell The key intuition for the definition is the classic notion of *but-for* causality used in the law literature: A is a cause of B if, but for A, B would not have happened. As is well known (see Example 2.2), but-for causality does not suffice to give a definition of causality. There are times we would like to call A a cause of B even if A is not a but-for cause of B. The definition of actual causality in causal models and our new abstract definition can both be viewed as essentially saying that A is a cause of B if, conditional on some appropriate C, A is a but-for cause of B. Like the definition of [7], we just require C to be a formula that is true (in the current state of the world).

In more detail, our notion says that A *is an actual cause of* B with respect to a language $\mathscr{L}$ if, roughly speaking, the following three conditions (which are formalized in Section 4) hold:

AC1′. A and B both hold (in the true state of the world);

AC2′. There is a formula $C \in \mathscr{L}$ such that C holds (in the true state of the world), and if $\neg A \wedge C$ were to hold (counterfactually) then $\neg B$ would hold.

AC3′. A is minimal; that is, there is no formula A' that implies A yet is not implied by A that also satisfies AC2′.

Note that the definition above is parametrized by a language $\mathscr{L}$. As we show, in case this language $\mathscr{L}$ consists only of conjunctions of primitive events (and A, B are also restricted to this language), then our definition collapses down to the HP definition of causality (see Theorem 4.2). Furthermore, if the language is slightly richer, then the same is true even in *recursive counterfactual structures*, counterfactual structures that essentially capture causal models (see Theorem 4.5).

Paper Outline The rest of the paper is organized as follows. In Section 2, we review the HP definition of actual causality, as modified by Halpern [7]. In Section 3, we discuss *causal-counterfactual families (of models)—ccfs*, a general class of frameworks to which our abstract definition applies. We show that causal models and the counterfactual structures considered in the philosophy literature are instances of ccfs. In Section 4, we give our abstract definition of causality, show that the HP definition is a special case of it, and consider its implications in counterfactual structures. We consider backtracking in Section 5, and explanation in Section **??**. Proofs are deferred to the appendix.

2 Actual causality in causal models

Here we briefly review the definition of causal models introduced by Halpern and Pearl [8], and the definition of actual causality in causal models. In fact, three variants of the definition have been proposed, called the “original”, “updated”, and “modified” definitions [7]. We focus here on the modified definition, both because it is the simplest to state and, as shown by Halpern [7], the easiest to work with, and because it lends itself naturally to generalization. The material in the first three subsections of this section is largely taken from [7]. While there are a number of other definitions of causality in causal models (e.g., [4, 5, 11, 12, 22]), we focus on the HP definition here because our generalization is based on it and because it has been the most influential (judging by Google scholar citations). We believe that our approach should be applicable to other ways of defining causality, but we have not checked details.

2.1 Causal models

We assume that the state is described in terms of variables and their values. Some variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. It is conceptually useful to split the variables into two sets: the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. The structural equations describe how the latter values are determined.

Formally, a *causal model* M is a pair $(\mathscr{S}, \mathscr{F})$, where $\mathscr{S}$ is a *signature*, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathscr{F}$ defines a set of *(modifiable) structural equations*, relating the values of the variables. A signature $\mathscr{S}$ is a tuple $(\mathscr{U}, \mathscr{V}, \mathscr{R})$, where $\mathscr{U}$ is a set of exogenous variables, $\mathscr{V}$ is a set of endogenous variables, and $\mathscr{R}$ associates with every variable $X \in \mathscr{U} \cup \mathscr{V}$ a nonempty set $\mathscr{R}(X)$ of possible values for X (i.e., the set of values over which X *ranges*). For simplicity, we assume here that $\mathscr{V}$ is finite, as is $\mathscr{R}(X)$ for every endogenous variable $X \in \mathscr{V}$. $\mathscr{F}$ associates with each endogenous variable $X \in \mathscr{V}$ a function denoted F_X (i.e., $F_X = \mathscr{F}(X)$) such that $F_X : (\times_{U \in \mathscr{U}} \mathscr{R}(U)) \times (\times_{Y \in \mathscr{V} - \{X\}} \mathscr{R}(Y)) \rightarrow \mathscr{R}(X)$. This mathematical notation just makes precise the fact that F_X determines the value of X, given the values of all the other variables in $\mathscr{U} \cup \mathscr{V}$. If there is one exogenous variable U and three endogenous variables, X, Y, and Z, then F_X defines the values of X in terms of the values of Y, Z, and U. For example, we might have $F_X(u,y,z) = u + y$, which is usually written as $X = U + Y$. Thus, if $Y = 3$ and $U = 2$, then $X = 5$, regardless of how Z is set.

The structural equations define what happens in the presence of external interventions. Setting the value of some variables $\vec{X}$ to $\vec{x}$ in a causal model $M = (\mathscr{S}, \mathscr{F})$ results in a new causal model, denoted $M_{\vec{X} \leftarrow \vec{x}}$, which is identical to M, except that, for each $X \in \vec{X}$, the equation in $\mathscr{F}$ for X is replaced by $X = x$, where x is the component of X in $\vec{x}$.

The dependencies between variables in a causal model $M = ((\mathscr{U}, \mathscr{V}, \mathscr{R}), \mathscr{F})$ can be described using a *causal network* (or *causal graph*), whose nodes are labeled by the endogenous and exogenous variables in M, with one node for each variable in $\mathscr{U} \cup \mathscr{V}$. The roots of the graph are (labeled by) the exogenous variables. There is a directed edge from variable X to Y if Y *depends on* X; this is the case if there is some setting of all the variables in $\mathscr{U} \cup \mathscr{V}$ other than X and Y such that varying the value of X in that setting results in a variation in the value of Y; that is, there is a setting $\vec{z}$ of the variables other than X and Y and values x and x' of X such that $F_Y(x, \vec{z}) \neq F_Y(x', \vec{z})$. A causal model M is *recursive* (or *acyclic*) if its causal graph is acyclic. If M is an acyclic causal model, then given a *context*, that is, a setting $\vec{u}$ for the exogenous variables in $\mathscr{U}$, the values of all the other variables are determined (i.e., there is a unique solution to all the equations). We can determine these values by starting at the top of the graph and working our way down. In this paper, following most of the literature, we restrict to recursive models.

2.2 Reasoning about causality

Given a signature $\mathscr{S} = (\mathscr{U}, \mathscr{V}, \mathscr{R})$, a *primitive event* is a formula of the form $X = x$, for $X \in \mathscr{V}$ and $x \in \mathscr{R}(X)$. A *basic causal formula (over $\mathscr{S}$)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$, where φ is a Boolean combination of primitive events, $Y_1, \ldots, Y_k$ are distinct variables in $\mathscr{V}$, and $y_i \in \mathscr{R}(Y_i)$. Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\varphi$. The special case where $k = 0$ is abbreviated as φ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$ says that φ would hold if Y_i were set to y_i, for $i = 1, \ldots, k$. A *causal formula* is a Boolean combination of basic causal formulas. Let $\mathscr{L}(\mathscr{S})$ consist of all causal formulas over the signature $\mathscr{S}$. (We typically omit $\mathscr{S}$ when it is clear from context.)

A pair $(M, \vec{u})$ consisting of a causal model M and a context $\vec{u}$ is a *(causal) setting*. A causal formula ψ is true or false in a setting. We write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in the setting $(M, \vec{u})$. The $\models$ relation is defined inductively. $(M, \vec{u}) \models X = x$ if the variable X has value x in the unique (since we are dealing with acyclic models) solution to the equations in M in context $\vec{u}$ (i.e., the unique vector of values for the variables that simultaneously satisfies all equations in M with the variables in $\mathscr{U}$ set to $\vec{u}$); $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y} = \vec{y}}, \vec{u}) \models \varphi$; and Boolean combinations are defined in the standard way.

2.3 The modified definition of actual causality

We are now in a position to state the (modified) definition of actual causality in causal models.

Definition 2.1: $\vec{X} = \vec{x}$ is an *actual cause of φ in the causal setting* $(M, \vec{u})$ if the following three conditions hold:

AC1. $(M,\vec{u}) \models (\vec{X} = \vec{x})$ and $(M,\vec{u}) \models \varphi$.

AC2. There is a set $\vec{W}$ of variables in $\mathcal{V}$ disjoint from $\vec{X}$, $\vec{w}^* \in \mathcal{R}(\vec{W})$, and $\vec{x} \in \mathcal{R}(\vec{X})$ such that $(M,\vec{u}) \models \vec{W} = \vec{w}^*$ and $(M,\vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}^*]\neg\varphi$.[1]

AC3. $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ satisfies condition AC2, where $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}$.[2]

∎

2.4 The role of AC2

The motivation for AC2 was to be able to deal with examples like the following, due to Lewis [15]. (The following analysis is largely taken from [7].)

Example 2.2: Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Because both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw.

We would like to call Suzy's throw the cause of the bottle shattering. This is the case in an appropriate causal model. The following causal model M^{rt} does the job. There are five endogenous variables:

- *ST* for "Suzy throws", with values 0 (Suzy does not throw) and 1 (she does);
- *BT* for "Billy throws", with values 0 (he doesn't) and 1 (he does);
- *BS* for "bottle shatters", with values 0 (it doesn't shatter) and 1 (it does).
- *BH* for "Billy's rock hits the (intact) bottle", with values 0 (it doesn't) and 1 (it does); and
- *SH* for "Suzy's rock hits the bottle", again with values 0 and 1.

For simplicity, assume that there is one exogenous variable U, with values ij, $i, j \in \{0,1\}$, which determines whether Billy and Suzy throw.

We consider the following equations for the variables:

- $ST = 1$ iff $U \in \{10,11\}$;
- $BT = 1$ iff $U \in \{01,11\}$;
- $SH = 1$ iff $ST = 1$;
- $BH = 1$ iff $BT = 1$ and $SH = 0$ (this builds into the equation the fact that Billy rock does not hit the intact bottle if Suzy's rock already hit it);
- $BS = 1$ iff $SH = 1$ or $BH = 1$.

Now consider the context $U = 11$, where both Suzy and Billy throw rocks. In this context, most people want to view $ST = 1$ as a cause of the bottle shattering, while $BT = 1$ is not. But $ST = 1$ is not a but-for cause of the bottle shattering; had Suzy not thrown, the bottle would have shattered anyway (Billy's rock would have hit it). Nevertheless, it is a cause. To show that $ST = 1$ is a cause, we take W in AC2 to be $\{BH\}$, and fix BH at its actual value of 0 (in fact, Billy 's throw does not hit the bottle). Note that $(M,11) \models [ST \leftarrow 0, BH \leftarrow 0]BS = 0$, so AC2 holds for $ST = 0$; clearly AC1 and AC3 hold as well. Thus, $ST = 1$ is a cause of $BS = 1$ in the context $U = 11$. It is easy to see that the symmetric argument does *not* show that $BT = 1$ is a cause. We cannot fix $SH = 0$, because in the actual state, SH is 1, not 0. We can think of $ST = 1$ as a but-for cause of $BS = 1$, conditional on $BH = 0$. ∎

---

[1] Although the definition in [7] does not explicitly say that $\vec{W}$ is disjoint from $\vec{X}$, since interventions are performed on distinct variables, this follows from the notation.

[2] Note that $\vec{X}' = \vec{x}'$ is guaranteed to satisfy AC1, which is why only AC2 is mentioned here.

The idea of fixing some variables at their actual values, as embodied by AC2, seems at first somewhat *ad hoc*. While it does give us the "right" answer in the example above, why should we want to do it in general? The idea of keeping some variables fixed at their actual value turns out to be quite a powerful one. This is perhaps best seen if we consider causality along a path. To take a simple example, suppose that someone's gender affects the outcome of a loan decision in two ways. First, gender affects salary, which in turns affects the decision. Second, gender affects how reliable someone is viewed as being (say, for example, that men are viewed as less likely to repay loans, even fixing everything else). That is, we have a simple causal graph with variables G (gender) that affects both S (salary) and R (reliability), while S and R both affect L (loan repayment).

We might decide that it is legitimate to take salary into account when considering loan repayment, but not the perception of reliability. That is, we might decide that it is legitimate to consider the effect of gender along the path $G-S-L$, but not along the path $G-R-L$. We can model causality along a path by simply fixing the values of variables at their actual values off the path. In this case, since we do not want to consider the impact of perception of reliability, we fix R at its actual value, which allows us to consider the impact of changing G along the path $G-S-L$.[3]

Fixing the values of some variables $\vec{W}$ is useful beyond being able to capture path causality. It allows us to consider causality conditional on certain features of the state being kept fixed—the key feature of our more abstract generalized definition.

3 Causal-counterfactual families

We want to define causality for all families of structures that satisfy certain minimal conditions. To that end, we consider (for lack of a better name) *causal-counterfactual families (ccfs)*. A ccf $\mathscr{M}$ is a family of models or structures (we use the latter two words more or less interchangeably) such that each model $M \in \mathscr{M}$ has a set of possible *states* or worlds, and there is a semantics that allows us to define $(M,w) \models \varphi$ for all formulas in a language $\mathscr{L}^{\mathscr{M}}$. We require that $\mathscr{L}^{\mathscr{M}}$ include primitive events in some signature $\mathscr{S}$, and be closed under conjunction and negation (so contains $\mathscr{L}(\mathscr{S})$) and *basic counterfactuals*, so that if φ and ψ are in $\mathscr{L}(\mathscr{S})$, then $\varphi \rightarrow \psi \in \mathscr{L}^{\mathscr{M}}$. But $\mathscr{L}^{\mathscr{M}}$ may also be far more expressive and include, for example, nested counterfactuals, and modalities such as belief and time. The richness of $\mathscr{L}^{\mathscr{M}}$ is exactly what will allow our abstract definition of causality to apply to quite expressive languages.

3.1 Causal models as a ccf

We can almost view causal models (over $\mathscr{S}$) as a ccf $\mathscr{M}$. The states are the contexts. Let the language $\mathscr{L}^+(\mathscr{S})$ be the minimal language allowed, namely, the result of starting with $\mathscr{L}(\mathscr{S})$ and then extending it to include formulas of the form $\varphi \rightarrow \psi$, where φ and ψ are in $\mathscr{L}(\mathscr{S})$. (Thus, $\mathscr{L}^+(\mathscr{S})$ has nested counterfactuals, while $\mathscr{L}(\mathscr{S})$ does not.) We take $\mathscr{L}^{\mathscr{M}} = \mathscr{L}^+(\mathscr{S})$.

We might hope to identify $\varphi \rightarrow \psi$ with $[\varphi]\psi$. This works fine as long as φ is a conjunction of primitive events. But $[\varphi]\psi$ is not defined in causal models for an arbitrary Boolean combination φ of primitive events. So, in order to view causal models as a ccf, we have to define the semantics of $\varphi \rightarrow \psi$ for an arbitrary Boolean combination φ of primitive events. To give semantics to a counterfactual of the form $\varphi \rightarrow \psi \in \mathscr{L}^+(\mathscr{S})$, we proceed as follows: Let $\vec{Y}$ be the endogenous variables that appear in φ. Then we take $(M,\vec{u}) \models \varphi \rightarrow \psi$ iff there exists a vector $\vec{y}$ of values such that $\varphi \wedge \vec{Y} = \vec{y}$ is consistent (when viewed as a propositional formula) and $(M,\vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\psi$. It is easy to check that $\varphi \rightarrow \psi$ agrees

[3] Thanks to Sander Beckers for pointing this out to us.

with $[\varphi]\psi$ if φ is a conjunction of primitive events, so this is a generalization of the standard definition of counterfactuals in causal models. This definition seems somewhat *ad hoc* and, indeed, has some counter-intuitive properties (see Example 4.4). For example, why restrict $\vec{Y}$ to consisting only of variables that appear in φ? Indeed, there are other definitions of counterfactuals in causal models that agree with $[\varphi]\psi$ if φ is a conjunction of primitive propositions. For example, we could take $\varphi \to \psi$ to be $[\varphi]\psi$ if φ is a conjunction of primitive propositions, and *false* otherwise. The current definition has the benefit of allowing us to prove that the generalized definition of causality agrees with the HP definition in causal models (see Theorem 4.2).

3.2 Counterfactual structures as a ccf

For our purposes, a (Lewis-style) *counterfactual structure* M over $\mathscr{S}$ is a tuple (S,R,π), where S is a finite set of states, π is an *interpretation* that maps each state to a truth assignment on the primitive events over $\mathscr{S}$, and R is a ternary relation over S. Intuitively, $(s,t,u) \in R$ if t is at least as close to s as u is.[4] Let $t \preceq_s u$ be an abbreviation for $(s,t,u) \in R$. Define $t \prec_s u$ if $t \preceq_s u$ and $u \not\preceq_s t$. For simplicity, we require that $\preceq_s$ be reflexive and transitive and assume that $s \prec_s t$ for all $t \neq s \in S$ (so, intuitively, s is the unique state closest to itself according to $\preceq_s$). These are standard assumptions in the literature.

As for the language, we again fix a signature $\mathscr{S}$. Taking $\mathscr{M}^c(\mathscr{S})$ to be the counterfactual structures over signature $\mathscr{S}$ (so that the interpretation π in these structures gives semantics to primitive events over $\mathscr{S}$), we assume that $\mathscr{L}^{\mathscr{M}^c(\mathscr{S})}$ includes $\mathscr{L}(\mathscr{S})$ and is closed under conjunction, negation, counterfactuals (so that if $\varphi,\psi \in \mathscr{L}^{\mathscr{M}^c(\mathscr{S})}$, then so is $\varphi \to \psi$, which means that we have nested counterfactuals). As usual, given a counterfactual structure $M \in \mathscr{M}^c(\mathscr{S})$ and state s, we define $(M,s) \models \varphi \to \psi$ if for all states t such that $(M,t) \models \varphi$ and there is no state t' such that $(M,t') \models \varphi$ and $t' \prec_s t$, we also have $(M,t) \models \psi$. That is, $(M,s) \models \varphi \to \psi$ if in all the closest states to s satisfying φ, ψ also holds. (It follows that $(M,s) \models \varphi \to \psi$ vacuously if there are no states t such that $(M,t) \models \varphi$.)

The careful reader will have noticed that this means we are treating formulas with negations differently in counterfactual structures and in causal models. For example, in the former case, the formula $(X \neq x) \to \psi$ is taken to be true at a state s if at the closest state to s such that $X \neq x$, ψ is true. In the latter case, the formula is true if, roughly speaking, there exists some value x' such that at the closest state where $X = x'$, ψ is true. If X is a binary variable and takes on only two values, say x and x', at the closest state where where $X \neq x$, we must have $X = x'$, and the two approaches agree. But if X is not binary, the two approaches may not agree. As we shall see, this distinction disappears given how we deal with causality.

4 A more general abstract definition of actual causality

We now define actual causality in arbitrary ccfs. While the focus of this paper is on causal models and counterfactual structures, we stress that the definition applies far more broadly. For example, rather than Lewis-style counterfactual structures, we could consider Stalnaker-style counterfactual structures [21], where there is a function f that associates with each state s and formula φ a unique state $f(s,\varphi)$ closest to s that satisfies φ. We can also apply the definition to subclasses of counterfactual structures (and do), alternate definitions of causal models, and the extended causal models of Lucas and Kemp [16]. In the definition, the set of possible witnesses is a parameter. As we shall see, this provides useful flexibility.

[4]The presentation here follows that of Halpern [6], with some simplifications for ease of exposition. Although we focus on Lewis's definition [14], largely the same approach handles that of Stalnaker [21], who assumes that there is a unique closest state, while Lewis allows for a set of closest states.

Definition 4.1: Given a ccf $\mathscr{M}$ and set $\mathscr{C}_{\mathscr{X}} \subseteq \mathscr{L}^{\mathscr{M}}$ of formulas, φ is an *actual cause of* ψ *with respect to* $\mathscr{C}_{\mathscr{X}}$ *in a model* $M \in \mathscr{M}$ *and state* s if the following three conditions hold:

AC1$'$. $(M,s) \models \varphi \wedge \psi$;

AC2$'$. There is a formula $\tau \in \mathscr{C}_{\mathscr{X}}$ such that $(M,s) \models \tau$ and $(M,s) \models (\neg\varphi \wedge \tau) \to \neg\psi$.

AC3$'$. φ is minimal; there is no formula $\varphi' \in \mathscr{C}_{\mathscr{X}}$ such that $\models \varphi \Rightarrow \varphi'$ and $\not\models \varphi' \Rightarrow \varphi$ such that AC2$'$ holds with φ replaced by φ'. ∎

4.1 Applying the abstract definition in actual causal models

This definition is, by design, quite similar to the definition of actual causality in causal models, but more general. Causes no longer have to have the form $\vec{X} = \vec{x}$, and we allow the witnesses in $\vec{W}$ to come from some arbitrary set $\mathscr{C}_{\vec{X}}$ of formulas that may depend on $\vec{X}$. But, as we observed above, there is a nontrivial difference in how negation is dealt with. As we hinted above, despite this difference, in causal models, we defined $\to$ so that AC2 and AC2$'$ lead to identical conclusions (with the appropriate choice of $\mathscr{C}_{\vec{X}}$).

Theorem 4.2: *Let $\mathscr{C}_{\vec{X}}$ consist of all formulas τ such that τ is a conjunction of arbitrary (non-negated) primitive events (so $\mathscr{C}_{\vec{X}}$ is in fact independent of $\vec{X}$). Then $\vec{X} = \vec{x}$ is a cause of ψ in a (recursive) causal setting $(M,\vec{u})$ according to AC1-3 iff $\vec{X} = \vec{x}$ is a cause of ψ with respect to $\mathscr{C}_{\vec{X}}$ in $(M,\vec{u})$ according to AC1$'$-3$'$.*

The choice of $\mathscr{C}_{\vec{X}}$ in Theorem 4.2 ensures that, just as in AC2, in whatever corresponds to the "closest states s' to s where $\vec{X} \neq \vec{x} \wedge \tau$ is true, a formula of the form $\vec{X} = x' \wedge \tau$ is true, where τ is the conjunction of all the primitive events in τ (but does not involve the negations of primitive events in τ). It easily follows that $(M,s) \models \tau$, which shows that AC2$'$ ends up with a formula much like that in AC2.

It is also worth noting that if $\mathscr{C}'_{\vec{X}} \supseteq \mathscr{C}_{\vec{X}}$, then the forward direction of Theorem 4.2 holds, but the converse may not. Theorem 4.2 does continue to hold if we allow negated primitive events as conjuncts, but once we allow disjunctions of primitive events, the converse fails in general. For example, if τ is $SH = 0 \vee BH = 0$ in Example 2.2, then clearly $(M^{rt}, U_{11}) \models \tau$. Moreover, $BT = 0 \wedge SH = 0 \wedge BH = 0$ is consistent with $BT \neq 1 \wedge \tau$, and $(M^{rt}, U_{11}) \models [BT \leftarrow 0, SH \leftarrow 0, BH \leftarrow 0](BS = 0)$, so with this choice of $\mathscr{C}'_{\mathscr{X}}$, $BT = 1$ would be a cause of $BS = 1$.

This example (and Theorem 4.2, for that matter) very much depend on the fact that $\vec{Y}$ consists precisely of the variables that appear on the left of the $\to$.

4.2 Applying the abstract definition of causality in recursive counterfactual structures

As we now show, the abstract definition of causality, when applied to a subfamily of counterfactual structures, can be viewed as generalizing the definition in causal structures. The subfamily of counterfactual structures corresponds in a precise sense to recursive causal models. Formally, a counterfactual structure M' *corresponds* to a causal model M if M and M' agree on all the equations; more precisely, for each endogenous variable Y, if $S_Y = (\mathscr{U} \cup \mathscr{V}) - \{Y\}$, then for each setting $\vec{s}_Y$ of the variables in S_Y, and each state s, in the closest states to s such that $S_Y = \vec{s}_Y$ holds, we have that $Y = F_Y(\vec{s}_Y)$. M' *strongly corresponds* to M if (a) M' corresponds to M, (b) for each assignment v of values to variables, there is a state in M' where v is the assignment, and (c) for each state s, if $(M',s) \models \mathscr{U} = \vec{u}$ and ψ is a propositional formula such that $\mathscr{U} = \vec{u} \wedge \psi$ is consistent, then in the closest states to s such that ψ holds, we have that

$\mathcal{U} = \vec{u}$.[5] Roughly speaking, $\mathcal{U} = \vec{u}$ continues to hold in the closest states to s unless some value in $\mathcal{U}$ is explicitly changed. (M', s) is (strongly) consistent with $(M, \vec{u})$ if M' (strongly) corresponds to M and $(M', s) \models \mathcal{U} = \vec{u}$.

A counterfactual structure is *recursive* if it strongly corresponds to a recursive causal model. In general, there may be many recursive counterfactual structures that strongly correspond to a given recursive causal model; the definition of strong consistency does not specify how the closest-state relation R in the counterfactual structure should work for formulas that are not structural equations. Nevertheless, as the following result shows, recursive counterfactual structures are closely related to (recursive) causal models, at least as far as counterfactual reasoning is concerned. We say that (M', s) is a *recursive setting* if M' is a recursive counterfactual structure and s is a state in M'.

Proposition 4.3: *[6, Theorem 3.4] If $(M, \vec{u})$ is a recursive causal setting and (M', s) is a recursive setting that is strongly consistent with $(M, \vec{u})$, then for all formulas $\varphi \in \mathcal{L}(\mathcal{S})$, $(M, \vec{u}) \models \varphi$ iff $(M', s) \models \varphi$ (where we identify $\psi \to \psi'$ with $[\psi]\psi'$).*

The restriction to $\mathcal{L}(\mathcal{S})$ is critical here, as the following example shows.

Example 4.4: Consider a recursive causal model M with two endogenous variables, X and Y, and one exogenous variable U, such that $\mathcal{R}(U) = \mathcal{R}(X) = \mathcal{R}(Y) = \{0, 1, 2\}$. Suppose that U is the unique parent of X, and X is the parent of Y; the equation for X is $X = U$, and the equation for Y is $Y = X$. Let ψ be the formula $(X \neq 0 \to Y = 1) \wedge (X \neq 0 \to Y = 2)$. Note that $\psi \notin \mathcal{L}(\mathcal{S})$ due to the antecedent $X \neq 0$. It is easy to see that $(M, U = 0) \models \psi$, but there can be no counterfactual structure M' and state s such that $(M', s) \models \psi$. ∎

Despite this example, we have the following result (whose proof is deferred to the appendix).

Theorem 4.5: *Let $\vec{X}$ be a set of endogenous variables, and let $\mathcal{C}_{\vec{X}}$ consist of all formulas τ such that τ is a conjunction of (a) arbitrary (non-negated) primitive events, and (b) a disjunction of the form $\vec{X} = \vec{x} \vee \vec{X} = \vec{x}^*$, where $\vec{x}^* \neq \vec{x}$. If $(M, \vec{u})$ is a recursive causal setting and (M', s) is a recursive counterfactual structure that is strongly consistent with $(M, \vec{u})$ then $\vec{X} = \vec{x}$ is a cause of ψ in $(M, \vec{u})$ according to AC1-3 iff $\vec{X} = \vec{x}$ is a cause of ψ with respect to $\mathcal{C}_{\vec{X}}$ in (M', s) according to AC1′-3′.*

Note that $\mathcal{C}_{\vec{X}}$ differs in Theorems 4.2 and 4.5. These results show the power of allowing $\mathcal{C}_{\vec{X}}$ as a parameter.

On the Role of the Language: We emphasize that for Theorem 4.5 to hold, we must allow the set $\mathcal{C}_{\vec{X}}$ of formulas that we condition on to contain (a quite restricted form of) disjunctions. If we were to allow arbitrary disjunctions, in particular, disjunctions that contain $\neg\psi$, then any formula would become a cause of ψ, and the definition becomes trivial. This issue can be dealt with by disallowing $\mathcal{C}_{\vec{X}}$ to contain formulas τ such that $\neg\varphi \wedge \tau$ implies ψ in *all* possible states, yet $\neg\varphi$ alone does not imply ψ in all possible states. (Indeed, the disjunction from Theorem 4.5 satisfies this condition if the model is sufficiently rich.)

An even easier, and arguably more natural, way of dealing with this issue is to simply restrict $\mathcal{C}_{\vec{X}}$ to contain only a conjunction of primitive events. Then, however, our definition and the HP definition would diverge: Suppose that Bob is placed in a room with a ticking bomb, which can be defused by entering a secret 100-digit combination. Bob can decide to run away, or try pushing in a number combination. If

[5] Although formulas of the form $\mathcal{U} = \vec{u}$ are not in the language $\mathcal{L}(\mathcal{S})$, for the purposes of this discussion, we assume that we have extended the language to include them, with the obvious semantics.

Bob runs away and the bomb explodes, should his action of running away be considered the cause of it? According to the HP definition, it would (since if Bob had just entered the right combination, then the bomb would be defused). According to our definition (and disallowing disjunctions and negations in $\mathscr{C}_{\vec{X}}$), it would depend on what number combination Bob would enter in the closest world where he does not run away. For instance, if Bob knows the right combination, then him running away would be considered a cause, but if he does not, then it would not. Arguably, the new definition (without disjunctions) gives the more reasonable answer in this situation. On the other hand, if the number combination is short (say, 1 digit), then arguably the HP definition would give the right answer whereas for ours it depends on what number combination Bob would enter in the closest world where he stays (and on how many trials he gets).

5 Backtracking

As has often been observed, when considering the effect of an intervention such as $X = x$ in a causal model, no variables "upstream" of X in the causal network (more precisely, no variables that are not descendants of X) are affected; only descendants of X can be affected. By way of contrast, even in recursive counterfactual structures, in the state(s) closest to (M,s) where $X = x$ is true, it may well be the case that ancestors of X have changed their value. Such backtracking is allowed in Lewis-style models, when considering closest states. Kügelgen et al. [13] introduced a whole mechanism for allowing backtracking counterfactuals in causal models that involved changing the values of exogenous variables. Note that changing the values of an exogenous variable is not viewed as an intervention; rather, it is assumed that the agent considering whether $\vec{X} = \vec{x}$ is a cause of φ is uncertain about the value of the exogenous variable, and is considering what would happen to φ if the value of the exogenous variable were different. Here we show that our abstract definition of causality allows backtracking in a natural way, without having to introduce a special mechanism for it. Indeed, we can move smoothly from allowing backtracking to forbidding it, just by fixing an appropriate set of variables.

It follows from Proposition 4.5 that we have no backtracking in recursive counterfactual structures that strongly correspond to causal models. But once we allow the context to change when considering causality, we can get backtracking. This is exactly what happens when we move from counterfactual structures that strongly correspond to a causal model to ones that just correspond to a causal model.

Example 5.1: Recall the causal setting $(M^{rt}, U = 11)$ for Example 2.2, the rock-throwing example. Let $M' = (S, R, \pi)$ be consistent with M^{rt}. Let s be a state such that (M', s) is consistent with $(M^{rt}, U = 11)$; in particular, $(M', s) \models ST = BT = SH = BS = 1 \wedge BH = 0$. Define R so that the closest state s' to s is such that $(M', s') \models U = 00$, so that $(M', s') \models ST = BT = SH = BH = BS = 0$. With these choices, M' does not strongly correspond to M. It is easy to see that $(M', s) \models BT = 0 \rightarrow BS = 0$, so $BT = 1$ is a cause of $BS = 1$ in (M', s). By allowing the context to change in the closest state to s, we get backtracking. ∎

By working with a counterfactual setting (M', s) that corresponds to a causal setting $(M, \vec{u})$, we can capture causality with and without backtracking in $(M, \vec{u})$, simply by adding the appropriate conjuncts to the formula τ in AC2$'$. For example, if we want capture why $\vec{X} = \vec{x}$ is a cause of φ without backtracking, we simply add the conjunct $\mathscr{U} = \vec{u}$ to ψ. If we want to allow some backtracking, we may want to fix the values of some ancestors of $\vec{X}$ to their actual value, without fixing $\mathscr{U}$. This makes it clear that whether or not we have backtracking depends on the set of formulas $\mathscr{C}_{\mathscr{X}}$ from which τ is drawn.

6 An Abstract Definition of Explanation

Our approach can be extended to deal with causal explanation, using the HP definition [7, 9]. We start by reviewing the HP definition of explanation; for more details and intuition, see [7]. As noted in the introduction, the HP definition of causality is relative to an agent and, specifically, what the agent knows. In causal models, an agent's knowledge is captured by the agent's *epistemic state* $\mathscr{K}$, where $\mathscr{K}$ is a set of contexts settings with a fixed causal model M.[6] Intuitively, the contexts in $\mathscr{K}$ are the ones that the agent considers possible, and reflects the agent's uncertainty regarding how the world works (represented by the equations involving context $\vec{u}$) and what is currently true (represented by the context $\vec{u}$). As is standard, we say that an agent *knows* φ if φ is true at all the settings in $\mathscr{K}$.

We next give the formal definition of explanation in causal models, and then give intuition for the clauses, particularly EX1(a).

Definition 6.1: *$\vec{X}=\vec{x}$ is an explanation of ψ in a model M relative to a set $\mathscr{K}$ of contexts in M* if the following conditions hold:

EX1(a). If $\vec{u} \in \mathscr{K}$ and $(M,\vec{u}) \models \vec{X}=\vec{x} \wedge \psi$, then there exists a conjunct $X=x$ of $\vec{X}=\vec{x}$ and a (possibly empty) conjunction $\vec{Y}=\vec{y}$ such that $X=x \wedge \vec{Y}=\vec{y}$ is a cause of ψ in $(M,\vec{u})$.

EX1(b). $(M,\vec{u}') \models [\vec{X} \leftarrow \vec{x}]\psi$ for all contexts $\vec{u}' \in \mathscr{K}$.

EX2. $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}'=\vec{x}'$ satisfies EX1(a) and EX1(b), where $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}$.

EX3. For some $\vec{u} \in \mathscr{K}$, we have that $(M,\vec{u})) \models \vec{X}=\vec{x} \wedge \psi$. (The agent considers possible a setting where the explanation and explanandum both hold.)

The explanation is *nontrivial* if it also satisfies

EX4. $(M,\vec{u}') \models \neg(\vec{X}=\vec{x}) \wedge \psi$ for some $\vec{u}' \in \mathscr{K}$ (The explanation is not already known given the observation of ψ.) ∎

The key part of the definition is EX1(b). Roughly speaking, it says that the explanation $\vec{X}=\vec{x}$ is a *sufficient cause* for ψ: for all settings that the agent considers possible, intervening to set $\vec{X}$ to $\vec{x}$ results in ψ. (See [7, Chapter 2.6] for a formal definition of sufficient cause.) EX2 and EX3 should be fairly clear. EX4 is meant to capture the intuition (discussed in the introduction) that what counts as an explanation depends on what the agent knows. In the formal model, this is captured by the set $\mathscr{K}$. That leaves EX1(a). Roughly speaking, it says that the explanation causes the explanandum. But there is a tension between EX1(a) and EX1(b) here: we may need to add conjuncts to the explanation to ensure that it suffices to make ψ true in all contexts (as required by EX1(b)). But these extra conjuncts may not be necessary to get causality in all contexts. At least one of the conjuncts of $\vec{X}=\vec{x}$ must be part of a cause of ψ, but the cause can include extra conjuncts. To understand why, it is perhaps best to look at an example.

Example 6.2: Consider a version of the rock-throwing example where now in the causal model M it is possible that (a) Suzy's rock gets to the bottle first, (b) Billy's rock gets to the bottle first, or (c) both rocks arrive simultaneously. We capture this using 12 contexts of the form u_{ijk} where, as before $i=1$ if Suzy

[6] In [7] there is also assumed to be a probability on $\mathscr{K}$; for simplicity, we ignore the probability here. Also, in [7], an agent may be uncertain about the model as well as the context. Assuming that the model is fixed, as we are doing here, is without loss of generality. If an agent is uncertain about the model, that uncertainty can be encoded into the context by having a special exogenous variable u^* that determines the model; that is, the value of u^* gives the model. Given a model M, the equations in M can be encoded using equations that hold when $u^*=M$.

throws and $j=1$ if Billy throws. If $k=1$ (resp., $k=2$, $k=3$) and both Billy and Suzy throw, then Suzy's rock gets to the bottle first (resp., Billy's rock gets to the bottle first; the rocks arrive simultaneously). In all cases, just one rock hitting the bottle suffices for the bottle to shatter.

Consider $\mathcal{K}_1 = \{u_{111}, u_{112}, u_{101}\}$. An explanation for the bottle shattering relative to $\mathcal{K}_1$ is $ST = 1 \wedge BT = 1$. Each of $ST=1$ and $BT=1$ satisfies EX1(b). But while $ST=1$ is a cause of the bottle shattering in $(M, \vec{u}_{111})$ and $(M, \vec{u}_{101})$, it is not a cause of the bottle shattering in $(M, \vec{u}_{111})$, while $BT=1$ is. On the other hand, $BT=1$ is not a cause of the bottle shattering in $(M, \vec{u}_{112})$. Thus, EX1(a) holds for $ST=1 \wedge BT=1$, taking $ST=1$ to be the relevant conjunct in $(M, \vec{u}_{111})$ and $(M, \vec{u}_{111})$ and $BT=1$ to be the relevant conjunct in $(M, \vec{u}_{111})$. (Note that we do not have to take u_{101} into account for EX1(a), because $(M, \vec{u}_{111}) \not\models ST=1 \wedge BT=1$.) It is easy to see that EX2, EX3, and EX4 hold in this case.

Next consider $\mathcal{K}_2 = \{u_{111}, u_{112}\}$. In this case, $ST=1 \wedge BT=1$ is an explanation of the bottle shattering relative to $\mathcal{K}_2$, but a trivial one; it was already known.

Finally, consider $\mathcal{K}_3 = \{u_{003}, u_{103}, u_{013}, u_{113}\}$. In this case, each of $ST=1$ and $BT=1$ is an explanation of the bottle shattering relative to $\mathcal{K}_3$. However, note that in the context (M, u_{113}), $ST=1$ is *not* a cause of the bottle shattering; if Suzy doesn't throw, then the bottle still shatters (since Billy's rock hit the bottle). However, $ST=1 \wedge BT=1$ is a cause, and EX1(a) allows adding a conjunct $Y=y$ (in this case, $BT=1$) to get a cause. Note that $ST=1 \wedge BT=1$ is not an explanation of the bottle shattering; it violates the minimality condition EX2. ∎

We now show how to translate each of these conditions to our more abstract framework. To start with, the obvious analogue of $\mathcal{K}$ is just a set of worlds in some model in a ccf.

Definition 6.3: Given a ccf $\mathscr{M}$ and set $\mathscr{C}_{\mathscr{X}} \subseteq \mathscr{L}^{\mathscr{M}}$ of formulas, φ is an *explanation of ψ with respect to $\mathscr{C}_{\mathscr{X}}$ in a model $M \in \mathscr{M}$ and state s relative to a set $\mathcal{K}$ of states in M* if the following three conditions hold:

EX1(a)′. If $s \in \mathcal{K}$ and $(M,s) \models \varphi \wedge \psi$, then there exist formulas $\tau_1, \tau_2 \in \mathscr{C}_{\mathscr{X}}$ such that (i) τ_1 is not valid, (ii) $\models \varphi \Rightarrow \tau_1$, (iii) $\models \tau_2 \Rightarrow \tau_1$, and (iv) τ_2 is a cause of ψ with respect to $\mathscr{C}_{\mathscr{X}}$ in (M,s) (according to Definition 4.1).

EX1(b)′. $(M,s') \models \varphi \rightarrow \psi$ for all states $s' \in \mathcal{K}$.

EX2′. φ is minimal; there is no formula φ' such that $\models \varphi \Rightarrow \varphi'$ and $\not\models \varphi' \Rightarrow \varphi$ such that EX1(a) and EX1(b) hold with φ replaced by φ'.

EX3′. For some $s \in \mathcal{K}$, we have that $(M,s) \models \varphi \wedge \psi$.

Again, the explanation is nontrivial if

EX4′. $(M,s') \models \neg\varphi \wedge \psi$ for some $s' \in \mathcal{K}$. ∎

All the primed conditions except for EX1(a)′ are fairly direct translations of their unprimed counterparts. For EX1(a)′, note that the effect of adding a conjunction $\vec{Y} = \vec{y}$ to $X = x$ in EX1(a) is to get a formula that implies $X = x$; that is why we have $\models \tau_2 \Rightarrow \tau_1$ in part (iii) or EX1(a)′. Similarly, the conjunct $X = x$ in EX1(a) (which corresponds to τ_1 in EX1(a)′) is implied by $\vec{X} = \vec{x}$ (which correspond to φ in EX1(a)′). To understand the requirement that τ_1 not be valid, note that if it were valid, then conditions (ii) and (iii) would hold vacuously, so (iv) would hold as long as we can find a formula that causes ψ with respect to $\mathscr{C}_{\mathscr{X}}$ in (M,s). We want τ_1 to have a nontrivial connection to φ. In EX1(a), this is accomplished by making it a conjunct of φ. In EX1(a)′, this is accomplished by having it be a nonvalid formula that is implied by φ.

In any case, we now get the analogue of Theorem 4.2. showing that the abstract definition of explanation generalizes the original HP definition.

Theorem 6.4: *Let $\mathscr{C}_{\vec{X}}$ consist of all formulas τ such that τ is a conjunction of arbitrary (non-negated) primitive events, and let $\mathscr{K}$ be a set of contexts in a causal model M. Then $\vec{X} = \vec{x}$ is an explanation of ψ in a causal setting $(M,\vec{u})$ relative to $\mathscr{K}$ according to EX1–3 (resp., EX1–4) iff $\vec{X} = \vec{x}$ is an explanation of ψ with respect to $\mathscr{C}_{\vec{X}}$ relative to $\mathscr{K}$ in $(M,\vec{u})$ according to EX1′-3′ (resp., EX1′-4′).*

We also get an analogue of Theorem 4.5, but to get this we need an even tighter connection between recursive counterfactual structures and recursive causal models. A recursive causal model M and a recursive counterfactual structure M' are *compatible* if for every context $\vec{u}$ in M there is a state s in M' such that $(M,\vec{u})$ is strongly consistent with (M,s), and for every state s in M' there is a context $\vec{u}$ in M such that $(M,\vec{u})$ is strongly consistent with (M,s). A set $\mathscr{K}$ of contexts in M and a set $\mathscr{K}'$ of states in M' are *compatible* if for every context $\vec{u}$ in $\mathscr{K}$ there is a state s in $\mathscr{K}'$ such that $(M,\vec{u})$ is strongly consistent with (M,s), and for every state s in $\mathscr{K}'$ there is a context $\vec{u}$ in $\mathscr{K}$ such that $(M,\vec{u})$ is strongly consistent with (M,s). Finally, a structure M is *acceptable* if (M,s) is acceptable for all states s in M.

Theorem 6.5: *Let $\vec{X}$ be a set of endogenous variables, and let $\mathscr{C}_{\vec{X}}$ consist of all formulas τ such that τ is a conjunction of (a) arbitrary (non-negated) primitive events, and (b) a disjunction of the form $\vec{X} = \vec{x} \vee \vec{X} = \vec{x}^*$, where $\vec{x}^* \neq \vec{x}$. Given a recursive causal model M and a recursive counterfactual structure M' such that M and M' are compatible, and compatible sets $\mathscr{K}$ of contexts in M and $\mathscr{K}'$ of states in M', if $(M,\vec{u})$ is a recursive causal setting and (M',s) is strongly consistent with $(M,\vec{u})$, then $\vec{X} = \vec{x}$ is an explanation of ψ relative to $\mathscr{K}$ in $(M,\vec{u})$ according to EX1–3 (resp., EX1-4) iff $\vec{X} = \vec{x}$ is an explanation of ψ with respect to $\mathscr{C}_{\vec{X}}$ relative to $\mathscr{K}'$ in (M',s) according to EX1′-3′ (resp., EX1′-4′).*

7 Conclusions

We have given an abstract definition of causality that applies to arbitrary ccfs. In particular, it applies to causal models (where it agrees with the standard HP definition) and to Lewis-style counterfactual structures (where it generalizes the HP definition). As we have observed, having such a general definition lets us apply the definition far more broadly, and to richer, more expressive languages.

Roughly speaking, the abstract definition says that A is a cause of B if, conditional on C, A is a but-for cause of B. We have restricted C to come from some set that we denoted $\mathscr{C}_{\vec{X}}$. We are currently trying to understand the impact of $\mathscr{C}_{\vec{X}}$ on the definition. While in causal models it sufficed to require that $\mathscr{C}_{\vec{X}}$ be a conjunction of primitive events to capture the HP notion of causality, in counterfactual structures we also required $\mathscr{C}_{\vec{X}}$ to contain (restricted) disjunctions.

What happens if we were to restrict $\mathscr{C}_{\vec{X}}$ to contain only a conjunction of primitive events? As discussed above, at this point, the two definitions would diverge: Recall again the scenario where Bob is placed in a room with a ticking bomb, which can be defused by entering a secret 100-digit combination. Bob can decide to run away, or try pushing in a number combination. If Bob runs away and the bomb explodes, should his action of running away be considered the cause of it? According to the HP definition, it would (since if Bob had just entered the right combination, then the bomb would be defused). According to our definition (and disallowing disjunctions and negations in $\mathscr{C}_{\vec{X}}$), it would depend on what number combination Bob would enter in the closest world where he does not run away. For instance, if Bob knows the right combination, then him running away would be considered a cause, but if he does not, then it would not. Arguably, the new definition (without disjunctions) gives the more reasonable answer in this situation.

We can, to some extent, deal with the problem by considering *normality* [7], where, intuitively, we want to consider $\vec{X} = \vec{x}$ a cause of φ only if the formula $\vec{X} = \vec{x}' \wedge \vec{W} = \vec{w}^*$ in AC2 is not abnormal. As argued in [7], this restriction leads to an arguably more reasonable notion of causality. We are currently exploring the possibility of capturing normality by allowing disjunctions in $\mathscr{C}_X$, but restricting all the disjuncts to be normal, in some appropriate sense, or by restricting to formulas using a small number of disjunctions. Again, this would show that by conditioning on the appropriate set of formulas, we can capture important intuitions.

A Proofs

For the reader's convenience, we repeat the statement of the results.

Theorem 4.2: *Let $\mathscr{C}_{\vec{X}}$ consist of all formulas τ such that τ is a conjunction of arbitrary (non-negated) primitive events (so $\mathscr{C}_{\vec{X}}$ is in fact independent of $\vec{X}$). Then $\vec{X}=\vec{x}$ is a cause of ψ in a causal setting $(M,\vec{u})$ according to AC1-3 iff $\vec{X}=\vec{x}$ is a cause of ψ with respect to $\mathscr{C}_{\vec{X}}$ in $(M,\vec{u})$ according to AC1′-3′.*

Proof: Suppose that $\vec{X}=\vec{x}$ is a cause of ψ in $(M,\vec{u})$ according to AC1-3. Clearly AC1′ and AC3′ are essentially equivalent to AC1 and AC3, so to show that $\vec{X}=\vec{x}$ is a cause of ψ with respect to $\mathscr{C}_{\vec{X}}$ according to AC1′-3′, it suffices to show that AC2′ holds.

Since AC2 holds, there exist a set $\vec{W}$ of endogenous variables, $\vec{w}^*\in\mathscr{R}(\vec{W})$, and $\vec{x}'\in\mathscr{R}(\vec{X})$ such that $(M,\vec{u})\models\vec{W}=\vec{w}^*$ and $(M,\vec{u})\models[\vec{X}\leftarrow\vec{x}',\vec{W}\leftarrow\vec{w}^*]\neg\psi$. We must have $\vec{x}'\neq\vec{x}$, since $(M,\vec{u})\models[\vec{X}\leftarrow\vec{x},\vec{W}\leftarrow\vec{w}^*]\psi$. Let τ be $\vec{W}=\vec{w}^*$. Clearly, $\tau\in\mathscr{C}_{\vec{X}}$; by assumption, $(M,\vec{u})\models\tau$. Moreover, $\vec{X}=\vec{x}'\wedge\vec{W}=\vec{w}^*$ is consistent with $\vec{X}\neq\vec{x}\wedge\tau$ (since $\vec{W}$ and $\vec{X}$ are disjoint). Let $\vec{Y}$ consist of the endogenous variables in $\neg(\vec{X}=x)\wedge\tau$; note that $\vec{Y}=\vec{W}\cup\vec{X}$. Choose $\vec{y}\in\mathscr{R}(\vec{Y})$ so that $\vec{Y}=\vec{y}$ is $\vec{X}=\vec{x}'\wedge\vec{W}=\vec{w}^*$. Thus, $\vec{X}\neq\vec{x}\wedge\tau\wedge\vec{Y}=\vec{y}$ is consistent. By assumption, $(M,\vec{u})\models[\vec{Y}=\vec{y}]\neg\psi$. By definition, $(M,\vec{u})\models(\vec{X}\neq\vec{x}\wedge\tau)\rightarrow\neg\psi$, so AC2′ holds, as desired.

For the converse, suppose that $\vec{X}=\vec{x}$ is a cause of φ with respect to $\mathscr{C}_{\vec{X}}$ in $(M,\vec{u})$ according to AC1′-3′. Again, it suffices to show that AC2 holds in $(M,\vec{u})$. Since AC2′ holds, there must exist a formula $\tau\in\mathscr{C}_{\vec{X}}$ such that $(M,\vec{u})\models\tau$ and $(M,\vec{u})\models(\vec{X}\neq\vec{x}\wedge\tau)\rightarrow\neg\psi$. Let $\vec{Y}$ be the endogenous variables that appear in $\vec{X}\neq\vec{x}\wedge\tau$. By definition, there must be a vector $\vec{y}$ of values such that $\tau\wedge(\vec{Y}=\vec{y})$ is consistent and $(M,\vec{u})\models[\vec{Y}=\vec{y}]\neg\psi$. $\vec{Y}$ must include $\vec{X}$. Let $\vec{x}'$ be the restriction of $\vec{y}$ to the variables in $\vec{X}$.

For each variable $Y\in\vec{Y}'=\vec{Y}-\vec{X}$, there is a value y such that $Y=y$ is a conjunct of τ; $Y=y$ must also be a conjunct of $\vec{Y}=\vec{y}$, since $\vec{Y}=\vec{y}$ is consistent with τ. Thus, $(M,\vec{u})\models Y=y$ and, more generally, if $\vec{y}'$ is the restriction of the values in $\vec{y}$ to the variables in $\vec{Y}'$, we have that $(M,\vec{u})\models\vec{Y}'=\vec{y}'$. It follows that AC2 holds, taking $\vec{W}=\vec{Y}'$ and $\vec{x}'$ as defined above. ∎

Theorem 4.5: *Let $\mathscr{C}_{\vec{X}}$ consist of all formulas τ such that τ is a conjunction of (a) arbitrary (non-negated) primitive events, and (b) a disjunction of the form $\vec{X}=\vec{x}\vee\vec{X}=\vec{x}^*$, where $\vec{x}^*\neq\vec{x}$. If $(M,\vec{u})$ is a recursive causal setting and (M',s) is strongly consistent with $(M,\vec{u})$, then $\vec{X}=\vec{x}$ is a cause of ψ in $(M,\vec{u})$ according to AC1-3 iff $\vec{X}=\vec{x}$ is a cause of ψ with respect to $\mathscr{C}_{\vec{X}}$ in (M',s) according to AC1′-3′.*

Proof: Suppose that $\vec{X}=\vec{x}$ is a cause of ψ in a causal setting $(M,\vec{u})$ (according to AC1-3), and (M',s) is a recursive counterfactual structure strongly consistent with $(M,\vec{u})$. We want to show that $\vec{X}=\vec{x}$ is a cause of ψ in (M',s). The argument is similar in spirit to the first half of the proof of Theorem 4.2.

Again, AC1′ and AC3′ are immediate; we need to show that AC2′ holds. Since AC2 holds, there is a set $\vec{W}$ of endogenous variables, a setting $\vec{w}^*$ such that $(M,\vec{u})\models\vec{W}=\vec{w}^*$, and a setting $\vec{x}'$ of the variables in $\vec{X}$ such that $(M,\vec{u})\models[\vec{X}\leftarrow\vec{x}',\vec{W}\leftarrow w^*]\neg\psi$. Let τ be $\vec{W}=\vec{w}^*\wedge(\vec{X}=\vec{x}\vee\vec{X}=\vec{x}')$. Clearly, $\tau\in\mathscr{C}_{\vec{X}}$ and $(M,\vec{u}))\models\tau$. Since (M',s) is strongly consistent with $(M,\vec{u})$, by Proposition 4.3, we have that $(M',s)\models\tau$. Moreover, $(M',s)\models(\vec{X}\neq\vec{x}\wedge\tau)\rightarrow\neg\psi$. The result follows.

For the converse, suppose that $\vec{X}=\vec{x}$ is a cause of ψ with respect to $\mathscr{C}_{\vec{X}}$ in (M',s). To show that $\vec{X}=\vec{x}$ is a cause of ψ in $(M,\vec{u})$, yet again, it suffices to show that AC2 holds. Since AC2$'$ holds, there must exist a formula $\tau \in \mathscr{C}_{\vec{X}}$ such that $(M',s) \models \tau$, and $(M',s) \models (\vec{X} \neq \vec{x} \wedge \tau) \rightarrow \neg\psi$. There is a conjunction of the form $\tau' \wedge \vec{X} = \vec{x}^*$ of non-negated primitive events that is equivalent to $\vec{X} \neq \vec{x} \wedge \tau$, where τ' does not mention any variable in $\vec{X}'$. Thus, $(M',s) \models (\vec{X} = \vec{x}^* \wedge \tau') \rightarrow \neg\psi$. By Proposition 4.3, it follows that $(M,\vec{u}) \models [\vec{X} = \vec{x}^* \wedge \tau']\neg\psi$. Since $(M',s) \models \tau$, we must have $(M',s) \models \tau'$, so by Proposition 4.3, $(M,\vec{u}) \models \tau'$. Thus, taking $\vec{W} = \vec{w}^*$ in AC2 to be τ', it follows that AC2 holds, as desired. ∎

Acknowledgments: Joe Halpern's work was suppprted in part by NSF grant FMitF-2319186, ARO grant W911NF-17-1-0061, MURI grant W911NF-19-1-0217 from the ARO, and a grant from Open Philanthropy. Rafael Pass's work was supported in part by NSF Award CNS 2149305, AFOSR Award FA9550-23-1-0387, AFOSR Award FA9550-24-1-0267 and ERC Advanced Grant KolmoCrypt - 101142322. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders.

References

[1] F. Barbero, K. Schulz, F. R. Velázquez-Quesada & K. Xie (2023): *Observing interventions: A logic for thinking about experiments*. Journal of Logic and Computation 33, pp. 1152–1185, doi:10.1093/logcom/exac011.

[2] R. Briggs (2012): *Interventionist counterfactuals*. Philosophical Studies 160, pp. 139–166, doi:10.1007/s11098-012-9908-5.

[3] P. Gärdenfors (1988): *Knowledge in Flux*. MIT Press, Cambridge, MA.

[4] C. Glymour & F. Wimberly (2007): *Actual causes and thought experiments*. In J. Campbell, M. O'Rourke & H. Silverstein, editors: Causation and Explanation, MIT Press, Cambridge, MA, pp. 43–67.

[5] N. Hall (2007): *Structural equations and causation*. Philosophical Studies 132, pp. 109–136, doi:10.1007/s11098-006-9057-9.

[6] J. Y. Halpern (2013): *From causal models to possible-worlds models of counterfactuals*. Review of Symbolic Logic 6(2), pp. 81–101, doi:10.1017/s1755020312000305.

[7] J. Y. Halpern (2016): *Actual Causality*. MIT Press, Cambridge, MA.

[8] J. Y. Halpern & J. Pearl (2005): *Causes and explanations: a structural-model approach. Part I: causes*. British Journal for Philosophy of Science 56(4), pp. 843–887, doi:10.1093/bjps/axi147.

[9] J. Y. Halpern & J. Pearl (2005): *Causes and explanations: a structural-model approach. Part II: explanations*. British Journal for Philosophy of Science 56(4), pp. 889–911.

[10] C. G. Hempel (1965): *Aspects of Scientific Explanation*. Free Press, New York.

[11] C. Hitchcock (2001): *The intransitivity of causation revealed in equations and graphs*. Journal of Philosophy XCVIII(6), pp. 273–299, doi:10.2307/2678432.

[12] C. Hitchcock (2007): *Prevention, preemption, and the principle of sufficient reason*. Philosophical Review 116, pp. 495–532, doi:10.1215/00318108-2007-012.

[13] J. von Kügelgen, A. Mohamed & S. Beckers (2023): *Backtracking counterfactuals*. In: Proc. 2nd Conference on Causal Learning and Reasoning, pp. 9584–9594. Available in *Proceedings of Machine Learning Research*, Vol 213, pp. 1–20, 2023.

[14] D. Lewis (1973): *Counterfactuals*. Harvard University Press, Cambridge, MA.

[15] D. Lewis (2000): *Causation as influence*. Journal of Philosophy XCVII(4), pp. 182–197, doi:10.2307/2678389.

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

[17] W. C. Salmon (1970): *Statistical explanation*. In R. Kolodny, editor: *The Nature and Function of Scientific Theories*, University of Pittsburgh Press, Pittsburgh, PA, pp. 173–231.

[18] W. C. Salmon (1984): *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton, N.J.

[19] C. Sartorio (2006): *Disjunctive causes*. *Journal of Philosophy* 103(10), pp. 521–538, doi:10.5840/jphil20061031029.

[20] M. Soloviev & J. Y. Halpern (2021): *Security properties as nested causal statements*. Available at http://arxiv.org/abs/2104.00872.

[21] R. C. Stalnaker (1968): *A theory of conditionals*. In N. Rescher, editor: *Studies in Logical Theory*, Blackwell, Oxford, U.K., pp. 98–112.

[22] J. Woodward (2003): *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford, U.K.