

Appropriate Causal Models and Stability of Causation

Joseph Y. Halpern*

Cornell University
halpern@cs.cornell.edu

Abstract

Causal models defined in terms of structural equations have proved to be quite a powerful way of representing knowledge regarding causality. However, a number of authors have given examples that seem to show that the Halpern-Pearl (HP) definition of causality (Halpern & Pearl 2005) gives intuitively unreasonable answers. Here it is shown that, for each of these examples, we can give two stories consistent with the description in the example, such that intuitions regarding causality are quite different for each story. By adding additional variables, we can disambiguate the stories. Moreover, in the resulting causal models, the HP definition of causality gives the intuitively correct answer. It is also shown that, by adding extra variables, a modification to the original HP definition made to deal with an example of Hopkins and Pearl (2003) may not be necessary. Given how much can be done by adding extra variables, there might be a concern that the notion of causality is somewhat unstable. Can adding extra variables in a “conservative” way (i.e., maintaining all the relations between the variables in the original model) cause the answer to the question “Is $X = x$ a cause of $Y = y$?” to alternate between “yes” and “no”? Here it is shown that adding an extra variable can change the answer from “yes” to “no”, but after that, it cannot change back to “yes”.

1 Introduction

Causal models defined in terms of structural equations have proved to be quite a powerful way of representing knowledge regarding causality. For example, they have been used to find causes of errors in software (Beer *et al.* 2012) and have been shown to be useful in predicting human attributions of responsibility (Gerstenberg & Lagnado 2010; Lagnado, Gerstenberg, & Zultan 2013). However, a number of authors (Glymour *et al.* 2010; Livengood 2013;

Spohn 2008; Weslake 2014) have given examples that seem to show that the Halpern-Pearl (HP) definition of causality (Halpern & Pearl 2005) gives intuitively unreasonable answers. One contribution of this paper is to show that these “problematic” examples can be dealt with in a relatively uniform way, by being a little more careful about the choice of causal model.

The need to choose the causal model carefully has been pointed out frequently (Blanchard & Schaffer 2013; Hall 2007; Halpern & Pearl 2005; Halpern & Hitchcock 2010; Hitchcock 2001; 2007). A causal model is characterized by the choice of variables, the equations relating them, and which variables we choose to make exogenous and endogenous (roughly speaking, which are the variables we choose to take as given and which we consider to be modifiable). Different choices of causal model for a given situation can lead to different conclusions regarding causality. The choices are, to some extent, subjective. While some suggestions have been made for good rules of thumb for choosing random variables (e.g., in (Halpern & Hitchcock 2010)), they are certainly not definitive. Moreover, the choice of variables may also depend in part on the variables that the modeler is aware of.

In this paper, I consider the choice of representation in more detail in four examples. I show that in all these examples, the model originally considered (which I call the “naive” model) does not correctly model all the relevant features of the situation. I argue that we can see this because, in all these cases, there is another story that can be told, also consistent with the naive model, for which we have quite different intuitions regarding causality. This suggests that a more careful model is needed to disambiguate the stories. In the first four cases, what turns out to arguably be the best way to do the disambiguation is to add (quite well motivated) extra variables, which, roughly speaking, capture the mechanism of causality. In the final example, what turns out to be most relevant is the decision as to which variables to make exogenous. Once we model things more carefully, the HP approach gives the expected answer in all cases.

As already observed by Halpern and Hitchcock (2014), adding extra variables also lets us deal with two other concerns that resulted in changes to the original HP definition. In Section 4, I consider an example due Hopkins and Pearl (2003) that motivated one of the changes. After showing

*Supported in part by NSF grants IIS-0911036 and CCF-1214844, AFOSR grant FA9550-08-1-0438 and by the DoD Multidisciplinary University Research Initiative (MURI) program administered by AFOSR under grant FA9550-12-1-0040. Thanks to Chris Hitchcock and Jonathan Livengood for interesting discussions and useful comments. I also thank Jonathan for a careful proofreading of the paper, which uncovered many typos. Finally, I think Thomas Blanchard for pointing out a serious problem in an earlier version of Theorem 5.1.
Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

how this example can be dealt with by adding an extra variable in a natural way (without modifying the original HP definition), I show that this approach generalizes: we can always add extra variables so as to get a model where the original HP definition can be used. In the full paper, I discuss an example due to Hiddleston (2005) that motivated the addition of normality considerations to the basic HP framework (see Section 2). Again, adding an extra variable deals with this example.

All these examples show that adding extra variables can result in a cause becoming a non-cause. Can adding variables also result in a non-cause becoming a cause? Of course, without constraints, this can easily happen. Adding extra variables can fundamentally change the model. However, as I show, if variables are added in a conservative way (so as to maintain all the relations between the variables in the original model), this cannot happen. If $X = x$ is not a cause of $Y = y$, then adding extra variables to the model cannot make $X = x$ a cause of $Y = y$.

The rest of this paper is organized as follows. In the next section, I review the HP definition (and the original definition) and its extension to deal with normality, as discussed in (Halpern & Hitchcock 2014). I discuss the five examples in Section 3. In Section 4, I discuss how adding extra variables can deal with the Hopkins-Pearl example and, more generally, can obviate the need to modify the original HP definition. I prove that non-causality is stable in Section 5. I conclude in Section 6 with some discussion of the implications of these results. Details of some proofs are left to the full paper (available at www.cs.cornell.edu/home/halpern/papers/causalmodeling.pdf).

2 Review

In this section, I briefly review the definitions of causal structures, the HP definition(s) of causality, and the extension that takes into account normality given by Halpern and Hitchcock. The exposition is largely taken from (Halpern 2008). The reader is encouraged to consult (Halpern & Pearl 2005), and (Halpern & Hitchcock 2014) for more details and intuition.

2.1 Causal structures

The HP approach assumes that the world is described in terms of random variables and their values. Some random 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 random 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. For example, in a voting scenario, we could have endogenous variables that describe what the voters actually do (i.e., which candidate they vote for), exogenous variables that describe the factors that determine how the voters vote, and a variable describing the outcome (who wins). The structural equations describe how the outcome is determined (e.g., majority rules).

Formally, a *causal model* M is a pair $(S, \mathcal{F})$, where S is a

signature, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathcal{F}$ defines a set of *modifiable structural equations*, relating the values of the variables. A signature S is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (that is, the set of values over which Y *ranges*). For simplicity, I assume here that $\mathcal{V}$ is finite, as is $\mathcal{R}(Y)$ for every endogenous variable $Y \in \mathcal{V}$. $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$. 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 $\mathcal{U} \cup \mathcal{V}$. 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 $X = U + Y$.¹ 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 variable X to x in a causal model $M = (S, \mathcal{F})$ results in a new causal model, denoted $M_{X=x}$, which is identical to M , except that the equation for X in $\mathcal{F}$ is replaced by $X = x$.

Following (Halpern & Pearl 2005), I restrict attention here to what are called *recursive* (or *acyclic*) models. This is the special case where there is some total ordering $\prec$ of the endogenous variables (the ones in $\mathcal{V}$) such that if $X \prec Y$, then X is independent of Y , that is, $F_X(\dots, y, \dots) = F_X(\dots, y', \dots)$ for all $y, y' \in \mathcal{R}(Y)$. Intuitively, if a theory is recursive, there is no feedback. If $X \prec Y$, then the value of X may affect the value of Y , but the value of Y cannot affect the value of X . It should be clear that if M is an acyclic causal model, then given a *context*, that is, a setting $\vec{u}$ for the exogenous variables in $\mathcal{U}$, there is a unique solution for all the equations. We simply solve for the variables in the order given by $\prec$. The value of the variables that come first in the order, that is, the variables X such that there is no variable Y such that $Y \prec X$, depend only on the exogenous variables, so their value is immediately determined by the values of the exogenous variables. The values of variables later in the order can be determined once we have determined the values of all the variables earlier in the order.

2.2 A language for reasoning about causality

To define causality carefully, it is useful to have a language to reason about causality. Given a signature $S = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a *primitive event* is a formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *causal formula* (over S) is one of the form $[Y_1 \leftarrow y_1, \dots, Y_k \leftarrow y_k] \varphi$, where φ is a Boolean combination of primitive events, $Y_1, \dots, Y_k$ are distinct variables in $\mathcal{V}$, and $y_i \in \mathcal{R}(Y_i)$. Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}] \varphi$. The special case where $k = 0$ is abbreviated as

¹The fact that X is assigned $U + Y$ (i.e., the value of X is the sum of the values of U and Y) does not imply that Y is assigned $X - U$; that is, $F_Y(U, X, Z) = X - U$ does not necessarily hold.

φ . Intuitively, $[Y_1 \leftarrow y_1, \dots, Y_k \leftarrow y_k]\varphi$ says that φ would hold if Y_i were set to y_i , for $i = 1, \dots, k$.

A causal formula ψ is true or false in a causal model, given a context. As usual, I write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in causal model M given context $\vec{u}$. 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 $\vec{u}$ (that is, the unique vector of values for the exogenous variables that simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\vec{u}$). The truth of conjunctions and negations is defined in the standard way. Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y}=\vec{y}}, \vec{u}) \models \varphi$. I write $M \models \varphi$ if $(M, \vec{u}) \models \varphi$ for all contexts $\vec{u}$.

2.3 The definition(s) of causality

The HP definition of causality, like many others, is based on counterfactuals. The idea is that A is a cause of B if, if A hadn't occurred (although it did), then B would not have occurred. But there are many examples showing that this naive definition will not quite work. To take just one example, consider the following story, due to Ned Hall and already discussed in (Halpern & Pearl 2005), from where the following version is taken.

Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since 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 say that Suzy's throw is a cause of the bottle shattering, and Billy's is not. But if Suzy hadn't thrown, Billy's rock would have hit the bottle and shattered it.

The HP definition of causality is intended to deal with this example, and many others.

Definition 2.1: $\vec{X} = \vec{x}$ is an *actual cause* of φ in $(M, \vec{u})$ 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 partition of $\mathcal{V}$ (the set of endogenous variables) into two subsets $\vec{Z}$ and $\vec{W}^2$ with $\vec{X} \subseteq \vec{Z}$ and a setting $\vec{x}'$ and $\vec{w}$ of the variables in $\vec{X}$ and $\vec{W}$, respectively, such that if $(M, \vec{u}) \models Z = z$ for all $Z \in \vec{Z}$ (i.e., z is the value of the random variable Z in the real world), then both of the following conditions hold:

(a) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\varphi$.

(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}]\varphi$ for all subsets $\vec{W}'$ of $\vec{W}$ and all subsets $\vec{Z}'$ of $\vec{Z}$, where I abuse notation and write $\vec{W}' \leftarrow \vec{w}$ to denote the assignment where the variables in $\vec{W}'$ get the same values as they would in the assignment $\vec{W} \leftarrow \vec{w}$, and similarly for $\vec{Z}' \leftarrow \vec{z}$.

²I occasionally use the vector notation $(\vec{Z}, \vec{W}, \text{etc.})$ to denote a set of variables if the order of the variables matters, which it does when we consider an assignment such as $\vec{W} \leftarrow \vec{w}$.

AC3. $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions AC1 and AC2.

The tuple $(\vec{W}, \vec{w}, \vec{x}')$ is said to be a *witness* to the fact that $\vec{X} = \vec{x}$ is a cause of φ .

AC1 just says that $\vec{X} = \vec{x}$ cannot be considered a cause of φ unless both $\vec{X} = \vec{x}$ and φ actually happen. AC3 is a minimality condition, which ensures that only those elements of the conjunction $\vec{X} = \vec{x}$ that are essential are considered part of a cause; inessential elements are pruned. Without AC3, if dropping a lit cigarette is a cause of a fire then so is dropping the cigarette and sneezing.

AC2 is the core of the definition. We can think of the variables in $\vec{Z}$ as making up the “causal path” from $\vec{X}$ to φ . Intuitively, changing the value of some variable in $\vec{X}$ results in changing the value(s) of some variable(s) in $\vec{Z}$, which results in the values of some other variable(s) in $\vec{Z}$ being changed, which finally results in the value of φ changing. The remaining endogenous variables, the ones in $\vec{W}$, are off to the side, so to speak, but may still have an indirect effect on what happens. AC2(a) is essentially the standard counterfactual definition of causality, but with a twist. If we want to show that $\vec{X} = \vec{x}$ is a cause of φ , we must show (in part) that if $\vec{X}$ had a different value, then so too would φ . However, this effect of the value of $\vec{X}$ on the value of φ may not hold in the actual context; essentially, it ensures that $\vec{X}$ alone suffices to bring about the change from φ to $\neg\varphi$; setting $\vec{W}$ to $\vec{w}$ merely eliminates possibly spurious side effects that may mask the effect of changing the value of $\vec{X}$. Moreover, although the values of variables on the causal path (i.e., the variables $\vec{Z}$) may be perturbed by the change to $\vec{W}$, this perturbation has no impact on the value of φ . As I said above, if $(M, \vec{u}) \models \vec{Z} = \vec{z}$, then z is the value of the variable Z in the context $\vec{u}$ (i.e., in the actual situation). We capture the fact that the perturbation has no impact on the value of φ by saying that if some variables $\vec{Z}$ on the causal path were set to their original values in the context $\vec{u}$, φ would still be true, as long as $\vec{X} = \vec{x}$. Roughly speaking, it says that if the variables in $\vec{X}$ are reset to their original value, then φ holds, even under the contingency $\vec{W}' = \vec{w}$ and even if some variables in $\vec{Z}$ are given their original values (i.e., the values in $\vec{z}$). The fact that AC2(b) must hold even if only a subset $\vec{W}'$ of the variables in $\vec{W}$ are set to their values in $\vec{w}$ (so that the variables in $\vec{W} - \vec{W}'$ essentially act as they do in the real world) and only a subset of the variables in $\vec{Z}$ are set to their values in the actual world says that we must have φ even if some things happen as they do in the actual world. See Sections 3.1 and 4 for further discussion of and intuition for AC2(b).

The original HP paper (Halpern & Pearl 2001) used a weaker version of AC2(b). Rather than requiring that $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}]\varphi$ for all subsets $\vec{W}'$ of $\vec{W}$, it was required to hold only for $\vec{W}$. That is, the following condition was used instead of AC2(b).

AC2(b') $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}]\varphi$ for all

subsets $\vec{Z}'$ of $\vec{Z}$.

The change from AC2(b') to AC2(b) may seem rather technical, but it has some nontrivial consequences. One of the contributions of this paper is to examine whether it is necessary; see Section 4 for details.

To deal with other problems in the HP definition, various authors have added the idea of *normality* to the definition. This can be done in a number of ways. I now briefly sketch one way that this can be done, following the approach in (Halpern & Hitchcock 2014).

Take a *world* (in a model M) to be a complete assignment of values to all the endogenous variables in M .³ We assume a partial preorder $\succeq$ on worlds, that is, a reflexive transitive relation.⁴ Intuitively, if $s \succeq s'$, then s is at least as normal, or typical, as s' . We can use normality in the definition of causality in two ways. Say that a world s is a *witness world* for $\vec{X} = \vec{x}$ being a cause of φ in $(M, \vec{u})$ if there is a witness $(\vec{W}, \vec{w}, \vec{x}')$ to $\vec{X} = \vec{x}$ being a cause of φ and $s = s_{\vec{X}=\vec{x}', \vec{W}=\vec{w}, \vec{u}}$, where $s_{\vec{X}=\vec{x}', \vec{W}=\vec{w}, \vec{u}}$ is the world that results by setting $\vec{X}$ to $\vec{x}'$ and $\vec{W}$ to $\vec{w}$ in context $\vec{u}$. We can then modify AC2(a) so as to require that we consider $\vec{X} = \vec{x}$ to be a cause of φ in $(M, \vec{u})$ only if the witness world s for $\vec{X} = \vec{x}$ being a cause is such that $s \succeq s_{\vec{u}}$, where $s_{\vec{u}}$ is the world determined by context $\vec{u}$. This says that, in determining causality, we consider only possibilities that result from altering atypical features of a world to make them more typical, rather than vice versa. This captures an observation made by Kahneman and Miller (1986) regarding human ascriptions of causality.

A somewhat more refined use of normality is to use it to “grade” causes. Say that s is a *best witness* for $\vec{X} = \vec{x}$ being a cause of φ if s is a witness world for $\vec{X} = \vec{x}$ being a cause of φ and there is no other witness world s' for $\vec{X} = \vec{x}$ being a cause of φ such that $s' \succ s$. (Note that there may be more than one best witness.) We can then grade candidate causes according to the normality of their best witnesses (without requiring that there must be a witness s such that $s \succeq s_{\vec{u}}$). Experimental evidence suggests that people are focusing on the cause with the best witness (according to their subjective ordering on worlds); see, e.g., (Cushman, Knobe, & Sinnott-Armstrong 2008; Hitchcock & Knobe 2009; Knobe & Fraser 2008).

3 The Examples

In this section, I consider examples due to Spohn (2008), Glymour et al. (2010), and Livengood (2013). I go through these examples in turn. I set the scene by considering the rock-throwing example mentioned above.

³In (Halpern & Hitchcock 2014), a world is defined as a complete assignment of values to the *exogenous* variables, but this is a typo.

⁴ $\succeq$ is not necessarily a partial order; in particular, it does not necessarily satisfy *antisymmetry* (i.e., $s \succeq s'$ and $s' \succeq s$ does not necessarily imply $s = s'$).

3.1 Throwing rocks at bottles

A naive model of the rock-throwing story just has binary random variables ST , BT , and BS (for “Suzy throws”, “Billy throws”, and “bottle shatters”, with the value of BS given by the equation $BS = ST \vee BT$: the bottle shatters if one of Suzy and Billy throws. Call this model M_{RT} . Let u be the context that results in $ST = BT = 1$: Suzy and Billy both throw. M_{RT} is described in Figure 1. (Although I have included the exogenous variable here, in later figures exogenous variables are omitted for ease of presentation.)

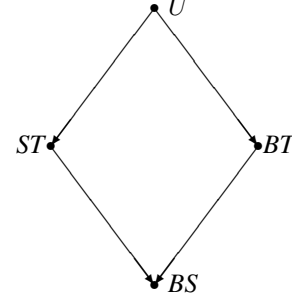


Figure 1: M_{RT} : the naive rock-throwing model.

As already pointed out by Halpern and Pearl (2005), in M_{RT} Suzy and Billy play completely symmetric roles. Not surprisingly, both $ST = 1$ and $BT = 1$ are causes of $BS = 1$ according to the HP definition. Clearly, M_{RT} cannot be used to distinguish a situation where Suzy is a cause from one where Billy is a cause.

In the story as given, people seem to agree that Suzy’s throw is a cause and Billy’s throw is not, since Suzy’s rock hit the bottle and Billy’s did not. M_{RT} does not capture this fact. Following Halpern and Pearl (2005), we extend M_{RT} so that it can express the fact that Suzy’s rock hit first by adding two more variables:

- 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.

The equations are such that $SH = ST$ (Suzy’s rock hits the bottle if Suzy throws), $BH = BT \wedge \neg SH$ (Billy’s rock hits an intact bottle if Billy throws and Suzy’s rock does not hit), and $BS = SH \vee BH$ (the bottle shatters if either Suzy’s rock or Billy’s rock hit it). Now if Suzy and Billy both throw ($ST = 1$ and $BT = 1$), Suzy’s rock hits the bottle ($SH = 1$), so that Billy’s rock does not hit an intact bottle ($BH = 0$). Call the resulting model M'_{RT} . M'_{RT} is described in Figure 2 (with the exogenous variable omitted).

In this model $BT = 1$ is *not* a cause. For example, if we take $\vec{Z} = \{BT, BH, BS\}$ in AC2 and set $ST = 0$, then while it is the case that $BS = 0$ if $BT = 0$ and $BS = 1$ if $BT = 1$, it is not the case that $BT = 1$ if we set BH to its original value of 0. This example shows the necessity of allowing some variables in $\vec{Z}$ to be set to their original values in AC2(b). This captures the intuition that Billy’s throw is not a cause because, in the actual world, his rock did not hit the bottle

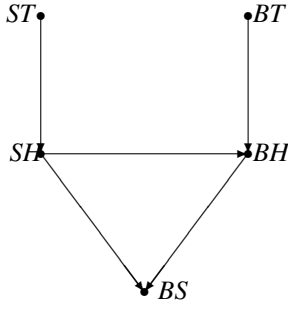


Figure 2: M'_{RT} : the better rock-throwing model.

($BH = 0$). By AC2(b), to establish $BT = 1$ as a cause of $BS = 1$, it would have to force $BS = 1$ even if $BH = 0$, which is not the case.

3.2 Spohn's example

The next example is due to Spohn (2008).

Example 3.1: There are four endogenous binary variables, A , B , C , and S , taking values 1 (on) and 0 (off). Intuitively, A and B are supposed to be alternative causes of C , and S acts as a switch. If $S = 0$, the causal route from A to C is active and that from B to C is dead; and if $S = 1$, the causal route from A to C is dead and the one from B to C is active. There are no causal relations between A , B , and S ; their values are determined by the context. The equation for C is $C = (\neg S \wedge A) \vee (S \wedge B)$.

Suppose that the context is such that $A = B = S = 1$, so $C = 1$. The HP definition yields $B = 1$ and $S = 1$ as causes of $C = 1$, as we would hope. But, unfortunately, it also yields $A = 1$ as a cause of $C = 1$. The argument is that in the contingency where S is set to 0, then if $A = 0$, then $C = 0$, while if $A = 1$, then $C = 1$. This does not seem so reasonable. Intuitively, if $S = 1$, then the value of A seems irrelevant to the outcome. Considerations of normality do not help here; all worlds seem to be equally normal.

But now consider a slightly different story. This time, we view B as the switch, rather than S . If $B = 1$, then $C = 1$ if either $A = 1$ or $S = 1$; if $B = 0$, then $C = 1$ only if $A = 1$ and $S = 0$. That is, $C = (B \wedge (A \vee S)) \vee (\neg B \wedge A \wedge \neg S)$. Although this is perhaps not as natural a story as the original, such a switch is surely implementable. In any case, a little playing with propositional logic shows that, in this story, C satisfies exactly the same equation as before: $(\neg S \wedge A) \vee (S \wedge B)$ is equivalent to $(B \wedge (A \vee S)) \vee (\neg B \wedge A \wedge \neg S)$. The key point is that, unlike the first story, in the second story, it seems to me quite reasonable to say that $A = 1$ is a cause of $C = 1$ (as are $S = 1$ and $B = 1$). Having $A = 1$ is necessary for the first “mechanism” to work.

Given that we have different causal intuitions for the stories, we should model them differently. One way to distinguish them is to add two more endogenous random variables, say D and E , that describe the ways that C could be 1. In Spohn's original story, we would have the equation $D = \neg S \wedge A$, $E = S \wedge B$, and $C = D \vee E$. In this model,

since $D = 0$ in the actual context, it is not hard to see that $A = 1$ is *not* a cause of $C = 1$, while $B = 1$ and $S = 1$ are, as they should be. Thus, in this model, we correctly capture our intuitions for the story.

To capture the second story, we can add variables D' and E' such that $D' = B \wedge (A \vee S)$, $E' = \neg B \wedge A \wedge S$, and $C = D' \vee E'$. In this model, it is not hard to see that all of $A = 1$, $B = 1$, and $S = 1$ are causes of $C = 1$.

This approach of adding extra variables leads to an obvious question: What is the role of these variables? I view D and E (resp., D' and E') as “structuring” variables, that help an agent “structure” a causal story. Consider Spohn's original story. We can certainly design a circuit where there is a source of power at A and B , a physical switch at S , and a bulb at C that turns on ($C = 1$) if either there is a battery at A ($A = 1$) and the switch is turned left ($S = 0$) or there is battery at B ($B = 1$) and the switch is turned right ($S = 1$). In this physical setup, there is no analogue of D and E . Nevertheless, to the extent that we view the models as a modeler's description of what is going, a modeler could usefully introduce D and E to describe the conditions under which $C = 1$, and to disambiguate this model from one where, conceptually, we might want to think of other ways that C could be 1 (as in the story with D' and E').

Note that we do not want to think of D as being *defined* to take the value 1 if $A = 1$ and $S = 0$. For then we could not intervene to set $D = 0$ if $A = 1$ and $S = 0$. Adding a variable to the model commits us to be able to intervene on it.⁵ In the real world, setting D to 0 despite having $A = 1$ and $S = 0$ might correspond to the connection being faulty when the switch is turned left. Indeed, since the equation for C is the same in both stories, it is only at the level of interventions that the difference between the two stories becomes meaningful. ■

3.3 Glymour et al.'s example

The next example is due to Glymour et al. (2010).

Example 3.2: A ranch has five individuals: $a_1, \dots, a_5$. They have to vote on two possible outcomes: staying around the campfire ($O = 0$) or going on a round-up ($O = 1$). Let A_i be the random variable denoting a_i 's vote, so $A_i = j$ if a_i votes for outcome j . There is a complicated rule for deciding on the outcome. If a_1 and a_2 agree (i.e., if $A_1 = A_2$), then that is the outcome. If $a_2, \dots, a_5$ agree, and a_1 votes differently, then then outcome is given by a_1 's vote (i.e., $O = A_1$). Otherwise, majority rules. In the actual situation, $A_1 = A_2 = 1$ and $A_3 = A_4 = A_5 = 0$, so by the first mechanism, $O = 1$. The question is what were the causes of $O = 1$.

Using the naive causal model with just the variables $A_1, \dots, A_5, O$, and the obvious equations describing O in terms of $A_1, \dots, A_5$, it is almost immediate that $A_1 = 1$ is a cause of $O = 1$. Changing A_1 to 0 results in $O = 0$. Somewhat surprisingly, in this naive model, $A_2 = 1$, $A_3 = 0$, $A_4 = 0$, and $A_5 = 0$ are also causes.⁶ To see that $A_2 = 1$

⁵I thank Chris Hitchcock for stressing this point.

⁶Glymour et al. point out that $A_1 = 1$, $A_3 = 0$, $A_4 = 0$, and $A_5 = 0$ are causes; they do not mention that $A_2 = 1$ is also a

is a cause, consider the contingency where $A_3 = 1$. Now if $A_2 = 0$, then $O = 0$ (majority rules); if $A_2 = 1$, then $O = 1$, since $A_1 = A_2 = 1$, and $O = 1$ even if A_3 is set back to its original value of 0. To see that $A_3 = 0$ is a cause, consider the contingency where $A_2 = 0$, so that all voters but a_1 vote for 0 (staying at the campsite). If $A_3 = 1$, then $O = 0$ (majority rules). If $A_3 = 0$, then $O = 1$, by the second mechanism (a_1 is the only vote for 0), while if A_2 is set to its original value of 1, then we still have $O = 1$, now by the second mechanism.

But all this talk of mechanisms (which is also implicit in Glymour et al. (2010); in footnote 11, they say that setting A_2 back to its original value of 1 “brings out the original result, but in a different way”) suggests that the mechanism should be part of the model. There are several ways of doing this. One is to add three new variables, call them M_1 , M_2 , and M_3 . These variables have values in $\{0, 1, 2\}$, where $M_j = 0$ if mechanism j is active and suggests an outcome 0, $M_j = 1$ if mechanism j is active and suggests an outcome of 1, and $M_j = 2$ if mechanism j is not active. (We actually don’t need the value $M_3 = 2$; mechanism 3 is always active, because there is always a majority with 5 voters all of whom must vote.) Note that at most one of the first two mechanisms can be active. We have obvious equations linking the value of M_1 , M_2 , and M_3 to the values of $A_1, \dots, A_5$.

Now the value of O just depends on the values of M_1 , M_2 , and M_3 : if $M_1 \neq 2$, then $O = M_1$; if $M_2 \neq 2$, then $O = M_2$, and if $M_1 = M_2 = 2$, then $O = M_3$. It is easy to see that in this model, if $A_1 = A_2 = 1$ and $A_3 = A_4 = A_5 = 0$, then none of $A_2 = 1$, $A_3 = 0$, $A_4 = 0$, and $A_5 = 0$ are causes. $A_1 = 1$ is cause, as we would expect, as is $M_2 = 1$, which seems reasonable: the second mechanism was the one that resulted in the outcome.

Now suppose that we change the description of the voting rule. We take $O = 1$ if one of the following two mechanisms applies:

- $A_1 = 1$ and it is not the case that both $A_2 = 0$ and exactly one of A_3 , A_4 , and A_5 is 1.
- $A_1 = 0$, $A_2 = 1$, and exactly two of A_3 , A_4 , and A_5 are 1.

It is not hard to check that, although the description is different, O satisfies the same equation in both stories. But now it does not seem so unreasonable that $A_2 = 1$, $A_3 = 0$, $A_4 = 0$, and $A_5 = 0$ are causes of $O = 1$. And indeed, if we construct a model in terms of these two mechanisms (i.e., add variables M'_1 and M'_2 that correspond to these two mechanisms), then it is not hard to see that $A_1 = 1$, $A_2 = 1$, $A_3 = 0$, $A_4 = 0$, and $A_5 = 0$ are all causes.

Here the role of the structuring variables M_1 , M_2 , and M_3 (resp. M'_1 and M'_2) as descriptors of the mechanism being invoked seems particularly clear. For example, setting $M_1 = 2$ says that the first mechanism will not be applied, even if $A_1 = A_2$; setting $M_1 = 1$ says that we act as if both a_1 and a_2 voted in favor, even if that is not the case. ■

cause.

3.4 Livengood’s voting examples

As Livengood (2013) points out, voting can lead to some apparently unreasonable causal outcomes (at least, if we model things naively). He first considers Jack and Jill, who live in an overwhelmingly Republican district. As expected, the Republican candidate wins with an overwhelming majority. Jill would normally have voted Democrat, but did not vote because she was disgusted by the process. Jack would normally have voted Republican, but did not vote because he (correctly) assumed that his vote would not affect the outcome. In the naive model, both Jack and Jill are causes of the Republican victory. For if enough of the people who voted Republican had switched to voting Democrat, then if Jack (or Jill) had voted Democrat, the Democrat would have won, while he would not have won had they abstained. Notice that, in this argument, Jack and Jill are treated the same way; their preferences make no difference.

We can easily construct a model that takes these preferences into account. One way to do so is to assume that their preferences are so strong that we may as well take them for granted. Thus, the preferences become exogenous; the only endogenous variables are whether or not they vote. In this case, Jack’s not voting is not a cause of the outcome, but Jill’s not voting is.

More generally, with this approach, a voter whose preference is made exogenous and is a strong supporter of the victor does not count as a cause of victory. This does not seem so unreasonable. After all, in an analysis of a close political victory in Congress, when an analyst talks about the cause(s) of victory, she points to the swing voters who voted one way or the other, not the voters that were taken to be staunch supporters of one particular side.

That said, making a variable exogenous seems like a somewhat draconian solution to the problem. It also does not allow us to take into account smaller gradations in depth of feeling. At what point should a preference switch from being endogenous to exogenous? We can achieve the same effect in an arguably more natural way by using normality considerations. In the case of Jack and Jill, we can take voting for a Democrat to be highly abnormal for Jack, and voting for a Republican to be highly abnormal for Jill. To show that either Jack (resp., Jill) is a cause of the victory, we need to consider a contingency where Jack (resp., Jill) votes for the Democratic candidate. This would be a change to a highly abnormal world in the case of Jack, but to a more normal world in the case of Jill. Thus, if we use normality as a criterion for determining causality, Jill would count as a cause, but Jack would not. If we use normality as a way of grading causes, Jack and Jill would still both count as causes for the victory, but Jill would be a much better cause. More generally, the more normal it would be for someone to vote Democratic, the better a cause that voter would be. The use of normality here allows for a more nuanced gradation of cause than the rather blunt approach of either making a variable exogenous or endogenous.

Now, following Livengood (2013), consider a vote where everyone can either vote for one of three candidates. Suppose that the actual vote is 17–2–0 (i.e., 17 vote for candidate A , 2 for candidate B , and none for candidate C). Then

not only is every vote for candidate A a cause of A winning, every vote for B is also a cause of A winning. To see this, consider a contingency where 8 of the voters for A switch to C . Then if one of the voters for B votes for C , the result is a tie; if that voter switches back to B , then A wins (even if some subset of the voters who switch from A to C switch back to A).

Is this reasonable? What makes it seem particularly unreasonable is that if it had just been a contest between A and B , with the vote 17–2, then the voters for B would not have been causes of A winning. Why should adding a third option make a difference?

In some cases it does seem reasonable that adding a third option makes a difference. For example, we speak of Nader costing Gore a victory over Bush in the 2000 election. But, as Livengood (2013) points out, we don’t speak of Gore costing Nader a victory, although in a naive HP model of the situation, all the voters for Gore are causes of Nader not winning as much as the voters for Nader are causes of Gore not winning. The discussion above points a way out of this dilemma. If a sufficiently large proportion of Bush and Gore voters are taken to be such strong supporters that they will never change their minds, and we make their votes exogenous, then it is still the case that Nader caused Gore to lose, but not the case that Gore caused Nader to lose. Similar considerations apply in the case of the 17–2 vote. (Again, we can use normality considerations to give arguably more natural models of these examples.)⁷

4 Do we need AC2(b)?

In this section, I consider the extent to which we can use AC2(b′) rather than AC2(b), and whether this is a good thing.

4.1 The Hopkins-Pearl example

I start by examining the Hopkins-Pearl example that was intended to show that AC2(b′) was inappropriate. The following description is taken from (Halpern & Pearl 2005).

Example 4.1: Suppose that a prisoner dies either if A loads B ’s gun and B shoots, or if C loads and shoots his gun. Taking D to represent the prisoner’s death and making the obvious assumptions about the meaning of the variables, we have that $D = (A \wedge B) \vee C$. Suppose that in the actual context u , A loads B ’s gun, B does not shoot, but C does load and shoot his gun, so that the prisoner dies. That is, $A = 1$, $B = 0$, and $C = 1$. Clearly $C = 1$ is a cause of $D = 1$. We would not want to say that $A = 1$ is a cause of $D = 1$, given that B did not shoot (i.e., given that $B = 0$).

⁷As a separate matter, most people would agree that Nader entering the race was a cause of Gore not winning, while Gore entering the race was not a cause of Nader not winning. Here the analysis is different. If Nader hadn’t entered, it seems reasonable to assume that there would have been no other strong third-party candidate, so just about all of Nader’s votes would have gone to Bush or Gore, with the majority going to Gore. On the other hand, if Gore hadn’t entered, there would have been another Democrat in the race replacing him, and most of Gore’s votes would have gone to the new Democrat in the race, rather than Nader.

However, suppose that we take the obvious model with the random variables A, B, C, D . With AC2(b′), $A = 1$ is a cause of $D = 1$. For we can take $\vec{W} = \{B, C\}$ and consider the contingency where $B = 1$ and $C = 0$. It is easy to check that AC2(a) and AC2(b′) hold for this contingency, so under the original HP definition, $A = 1$ is a cause of $D = 1$. However, AC2(b) fails in this case, since $(M, u) \models [A \leftarrow 1, C \leftarrow 0](D = 0)$. The key point is that AC2(b) says that for $A = 1$ to be a cause of $D = 1$, it must be the case that $D = 0$ if only some of the values in $\vec{W}$ are set to $\vec{w}$. That means that the other variables get the same value as they do in the actual context; in this case, by setting only A to 1 and leaving B unset, B takes on its original value of 0, in which case $D = 0$. AC2(b′) does not consider this case.

Nevertheless, as pointed out by Halpern and Hitchcock (2014), we can use AC2(b′) if we have the “right” model. Suppose that we add a new variable E such that $E = A \wedge B$, so that $E = 1$ iff $A = B = 1$, and set $D = E \vee C$. Thus, we have captured the intuition that there are two ways that the prisoner dies. Either C shoots, or A loads and B fires (which is captured by E). It is easy to see that (using either AC2(b) or AC2(b′)) $B = 0$ is *not* a cause of $D = 1$. ■

As I now show, the ideas of this example generalize. But before doing that, I define the notion of a conservative extension.

4.2 Conservative extensions

In the rock-throwing example, adding the extra variables converted $BT = 1$ from being a cause to not being a cause of $BS = 1$. Similarly, adding extra variables affected causality in all the other examples above. Of course, without any constraints, it is easy to add variables to get any desired result. For example, consider the rock-throwing model M'_{RT} . Suppose that we add a variable BH_1 with equations that set $BH_1 = BT$ and $BS = SH \vee BT \vee BH_1$. This results in a new “causal path” from BT to BS going through BH_1 , independent of all other paths. Not surprisingly, in this model, $BT = 1$ is indeed a cause of $BS = 1$.

But this seems like cheating. What we want is a model where the new variable(s) are added in a way that extends what we know about the old variables. Intuitively, suppose that we had a better magnifying glass and could look more carefully at the model. We might discover new variables that were previously hidden. But we want it to be the case that any setting of the old variables results in the same observations. This is made precise in the following definition.

Definition 4.2: A causal model $M' = ((\mathcal{U}', \mathcal{V}', \mathcal{R}'), \mathcal{F}')$ is a *conservative extension* of $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F})$ if $\mathcal{U} = \mathcal{U}'$, $\mathcal{V} \subseteq \mathcal{V}'$ if, for all contexts $\vec{u}$, all variables $X \in \mathcal{V}$, and all settings $\vec{w}$ of the variables in $\vec{W} = \mathcal{V} - \{X\}$, we have $(M, \vec{u}) \models [\vec{W} \leftarrow \vec{w}](X = x)$ iff $(M', \vec{u}) \models [\vec{W} \leftarrow \vec{w}](X = x)$. That is, no matter how we set the variables other than X , X has the same value in context $\vec{u}$ in both M and M' . ■

The following lemma gives further insight into the notion of conservative extension. It says that if M' is a conservative extension of M , then the same causal formulas involving only the variables in $\mathcal{V}$ hold in M and M' .

Lemma 4.3: Suppose that M' is a conservative extension of $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F})$. Then for all causal formulas φ that mention only variables in $\mathcal{V}$ and all context u , we have $(M, u) \models \varphi$ iff $(M', u) \models \varphi$.

Proof: See the full paper. ■

4.3 Avoiding AC2(b)

I now show that we can always use AC2(b') instead of AC2(b), if we add extra variables.

Theorem 4.4: If $X = x$ is not a cause of $Y = y$ in $(M, \vec{u})$ using AC2(b), but is a cause using AC2(b'), then there is a model M' that is a conservative extension of M such that $X = x$ is not a cause of $Y = y$ using AC2(b').

Proof: Suppose that $(\vec{W}, \vec{w}, x')$ is a witness to $X = x$ being a cause of $Y = y$ in $(M, \vec{u})$ using AC2(b'). Let $(M, \vec{u}) \models \vec{W} = \vec{w}$. We must have $\vec{w} \neq \vec{w}^*$, for otherwise it is easy to see that $X = x$ would be a cause of $Y = y$ in $(M, \vec{u})$ using AC2(b) with witness $(\vec{W}, \vec{w}, x')$.

If M' is a conservative extension of M with additional variables $\mathcal{V}'$, and $X = x$ is a cause of $Y = y$ in $(M, \vec{u})$ with witness $(\vec{W}, \vec{w}, x')$, we say that $(\vec{W}', \vec{w}', x')$ extends $(\vec{W}, \vec{w}, x')$, if $\vec{W} \subseteq \vec{W}' \subseteq \vec{W} \cup \mathcal{V}'$ and $\vec{w}'$ agrees with $\vec{w}$ on the variables in $\vec{W}$.

I now construct a conservative extension M' of M in which $X = x$ is not a cause of $Y = y$ using AC2(b') with a witness extending $(\vec{W}', \vec{w}, x')$. Of course, this just kills one witness. I then show that we can construct further extensions to kill all other witnesses to $X = x$ being a cause of $Y = y$ using AC2(b').

Let M' be obtained from M by adding one new variable N . All the variables have the same equations in M and M' except for Y and (of course) N . The equations for N are easy to explain: if $X = x$ and $\vec{W} = \vec{w}$, then $N = 1$; otherwise, $N = 0$. The equations for Y are the same in M and M' (and do not depend on the value of N) except for two special cases. To define these cases, for each variable $Z \in \mathcal{V} - \vec{W}$, if $x'' \in \{x, x'\}$ define $z_{x'', \vec{w}}$ as the value such that $(M, \vec{u}) \models [X \leftarrow x'', \vec{W} \leftarrow \vec{w}](Z = z_{x'', \vec{w}})$. That is, $z_{x'', \vec{w}}$ is the value taken by Z if X is set to x'' and $\vec{W}$ is set to $\vec{w}$. Let $\vec{V}'$ consist of all variables in $\mathcal{V}$ other than Y , let $\vec{v}'$ be a setting of the variables in $\vec{V}'$, and let $\vec{Z}'$ consist of all variables in $\vec{V}' - \vec{W}$ other than X . Then we want the equations for Y in M' to be such that for all $j \in \{0, 1\}$, we have

$$\begin{aligned} (M, \vec{u}) \models [\vec{V}' \leftarrow \vec{v}'](Y = y'') \text{ iff} \\ (M', \vec{u}) \models [\vec{V}' \leftarrow \vec{v}'; N \leftarrow j](Y = y'') \end{aligned}$$

unless the assignment $\vec{V}' \leftarrow \vec{v}'$ results in either (a) $X = x$, $\vec{W} = \vec{w}$, $Z = z_{x, \vec{w}}$ for all $Z \in \vec{Z}'$, and $N = 0$ or (b) $X = x'$, $\vec{W} = \vec{w}$, $Z = z_{x', \vec{w}}$ for all $Z \in \vec{Z}'$, and $N = 1$. (Note that in both of these cases, the value of N is “abnormal”. If $X = x$, $\vec{W} = \vec{w}$ and $Z = z_{x, \vec{w}}$ for all $z \in \vec{Z}'$, then N should be 1; if we set X to x' and change the values of the variables

in $\vec{Z}'$ accordingly, then N should be 0.) If (a) holds, $Y = y'$ in M' ; if (b) holds, $Y = y$.

Lemma 4.5:

- (a) It is not the case that $X = x$ is a cause of $Y = y$ using AC2(b') in M' with a witness that extends $(\vec{W}, \vec{w}, x')$.
- (b) M' is a conservative extension of M .
- (c) if $X = x$ is a cause of $Y = y$ in $(M', \vec{u})$ using AC2(b) (resp. AC2(b')) with a witness extending $(\vec{W}', \vec{w}', x'')$ then $X = x$ is a cause of $Y = y$ in $(M, \vec{u})$ using AC2(b) (resp. AC2(b')) with witness $(\vec{W}', \vec{w}', x'')$.

The proof of Lemma 4.5 and the remainder of the proof of Theorem 4.4 can be found in the full paper. ■

4.4 Discussion

Theorem 4.4 suggests that, by adding extra variables appropriately, we can go back to the definition of causality using AC2(b') rather than AC2(b). This has some technical advantages. For example, with AC2(b'), causes are always single conjuncts (Eiter & Lukasiewicz 2002; Hopkins 2001). As shown in (Halpern 2008), this is not in general the case with AC2(b); it may be that $X_1 = x_1 \wedge X_2 = x_2$ is a cause of $Y = y$ with neither $X_1 = x_1$ nor $X_2 = x_2$ being causes. It also seems that testing for causality is harder using AC2(b). Eiter and Lukasiewicz (2002) show that, using AC2(b'), testing for causality is NP-complete for binary models (where all random variables are binary) and Σ_2 -complete in general; with AC2(b), it seems to be Σ_2 -complete in the binary case and Π_3 -complete in the general case (Aleksandrowicz et al. 2013).

On the other hand, adding extra variables may not always be a natural thing to do. For example, in Beer et al.'s (2012) analysis of software errors, the variables chosen for the analysis are determined by the program specification. Moreover, Beer et al. give examples where AC2(b) is needed to get the intuitively correct answer. Unless we are given a principled way of adding extra variables so as to be able to always use AC2(b'), it is not clear how to automate an analysis. There does not always seem to be a “natural” way of adding extra variables so that AC2(b') suffices (even assuming that we can agree on what “natural” means!).

Adding extra variables also has an impact on complexity. Note that, in the worst case, we may have to add an extra variable for each pair (W, w) such that there is a witness (W, w, x') for $X = x$ being a cause of $Y = y$. In all the standard examples, there are very few witnesses (typically 1–2), but I have been unable to prove a nontrivial bound on the number of witnesses.

More experience is needed to determine which of AC2(b) and AC2(b') is most appropriate. Fortunately, in many cases, the causality judgment is independent of which we use.

5 The Stability on Non-Causality

The examples in Section 3 raises a potential concern. Consider the rock-throwing example again. Adding extra variables changed $BT = 1$ from being a cause of $BS = 1$ to

not being a cause. Could adding even more variables convert $BT = 1$ back to being a cause? Could it then alternate further?

As I now show, this cannot happen. Although, as we saw with the rock-throwing example, adding variables can convert a single-variable cause to a non-cause, it cannot convert a single-variable non-cause to a cause. Non-causality is stable.

Theorem 5.1: *If $X = x$ is not a cause of φ in (M, u) , and M' is a conservative extension of M , then $X = x$ is not a cause of φ in (M', u) .*

Proof: Suppose, by way of contradiction, that $X = x$ is not a cause of φ in (M, u) , M' is a conservative extension of M , and $\vec{X} = \vec{x}$ is a cause of φ in (M', u) with witness $(\vec{W}, \vec{w}, \vec{x}')$. Let $\vec{W}_1$ be the intersection of $\vec{W}$ with $\mathcal{V}$, the set of endogenous variables in M , let $\vec{Z}_1 = \mathcal{V} = \vec{W}$, and let $\vec{w}_1$ be the restriction of $\vec{w}$ to the variables in $\vec{W}_1$. Since $X = x$ is not a cause of φ in (M, u) , it is certainly not a cause with witness $(\vec{W}_1, \vec{w}_1, x')$, so either (a) $(M, u) \models X \neq x \vee \neg\varphi$; (b) $(M, u) \models [X \leftarrow x', \vec{W}_1 \leftarrow \vec{w}_1]\varphi$, or (c) there must be subsets $\vec{W}'_1$ of $\vec{W}_1$ and $\vec{Z}'_1$ of $\vec{Z}_1$ such that if $(M, u) \models \vec{Z}'_1 = \vec{z}_1$ (i.e., $\vec{z}_1$ gives the actual values of the variables in $\vec{Z}'_1$), then $(M, u) \not\models [X \leftarrow x, \vec{W}'_1 \leftarrow \vec{w}'_1, \vec{Z}'_1 \leftarrow \vec{z}_1]\neg\varphi$. Since M' is a conservative extension of M , by Lemma 4.3, if (a) holds, then $(M', u) \models X \neq x \vee \neg\varphi$; if (b) holds, $(M', u) \models [X \leftarrow x', \vec{W}_1 \leftarrow \vec{w}_1]\varphi$; and if (c) holds, then $(M', u) \not\models [X \leftarrow x, \vec{W}'_1 \leftarrow \vec{w}'_1, \vec{Z}'_1 \leftarrow \vec{z}_1]\neg\varphi$. Moreover, $(M', u) \models \vec{Z}'_1 = \vec{z}_1$. Thus, it follows that $X = x$ is not a cause of φ with witness $(\vec{W}, \vec{w}, x')$ in (M', u) . For if (a) holds, then AC1 fails; if (b) holds, then AC2(a) fails, and if (c) holds, then AC2(b) fails. ■

Notice that this argument does not quite work if the cause involves multiple variables. If $\vec{X} = \vec{x}$ is a cause of φ in (M', u) with witness $(\vec{W}, \vec{w}, \vec{x}')$, it is possible that it is not a cause in (M, u) because AC3 fails. While I believe that stability of non-causality holds even for causes involving multiple variables, proving it remains an open problem.

It is interesting that, in the single-variable case, non-causality is stable, while causality is not. Intuitively, this says that if $X = x$ is not a cause of $Y = y$, this is somehow encoded in the variables. For example, the fact that Billy's throw is not a cause of the bottle shattering in M'_{RT} is due to the variables BH and SH encoding the fact that Suzy's rock hits and Billy's does not. This information is maintained in any extension of M'_{RT} , so Billy's throw cannot become a cause again.

6 Conclusions

This paper has demonstrated the HP definition of causality is remarkably resilient, but it emphasizes how sensitive knowledge representation can be to the choice of models. The focus has been on showing that the choice of variables is a powerful modeling tool. But it is one that can be abused.

One lesson that comes out clearly is the need to have variables that describe the mechanism of causality, particularly if there is more than one mechanism. However, this is hardly a general recipe. Rather, it is a heuristic for constructing a “good” model. As Hitchcock and I pointed out (Halpern & Hitchcock 2010), constructing a good model is still more of an art than a science.

References

- Aleksandrowicz, G.; Chockler, H.; Halpern, J. Y.; and Ivrii, A. 2013. Complexity of causality: updated results. unpublished manuscript.
- Beer, I.; Ben-David, S.; Chockler, H.; Orni, A.; and Treffer, R. J. 2012. Explaining counterexamples using causality. *Formal Methods in System Design* 40(1):20–40.
- Blanchard, T., and Schaffer, J. 2013. Cause without default. unpublished manuscript.
- Cushman, F.; Knobe, J.; and Sinnott-Armstrong, W. 2008. Moral appraisals affect doing/allowing judgments. *Cognition* 108(1):281–289.
- Eiter, T., and Lukasiewicz, T. 2002. Complexity results for structure-based causality. *Artificial Intelligence* 142(1):53–89.
- Gerstenberg, T., and Lagnado, D. 2010. Spreading the blame: the allocation of responsibility amongst multiple agents. *Cognition* 115:166–171.
- Glymour, C.; Danks, D.; Glymour, B.; Eberhardt, F.; Ramsey, J.; Scheines, R.; Spirtes, P.; Teng, C. M.; and Zhang, J. 2010. Actual causation: a stone soup essay. *Synthese* 175:169–192.
- Hall, N. 2007. Structural equations and causation. *Philosophical Studies* 132:109–136.
- Halpern, J. Y., and Hitchcock, C. 2010. Actual causation and the art of modeling. In Dechter, R.; Geffner, H.; and Halpern, J., eds., *Causality, Probability, and Heuristics: A Tribute to Judea Pearl*. London: College Publications. 383–406.
- Halpern, J. Y., and Hitchcock, C. 2014. Graded causation and defaults. *The British Journal for the Philosophy of Science*. Available at <http://www.cs.cornell.edu/home/halpern/papers/normality.pdf>.
- Halpern, J. Y., and Pearl, J. 2001. Causes and explanations: A structural-model approach. Part I: Causes. In *Proc. Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI 2001)*, 194–202.
- Halpern, J. Y., and Pearl, J. 2005. Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for Philosophy of Science* 56(4):843–887.
- Halpern, J. Y. 2008. Defaults and normality in causal structures. In *Principles of Knowledge Representation and Reasoning: Proc. Eleventh International Conference (KR '08)*. 198–208.
- Hiddleston, E. 2005. Causal powers. *British Journal for Philosophy of Science* 56:27–59.
- Hitchcock, C., and Knobe, J. 2009. Cause and norm. *Journal of Philosophy* 106:587–612.

- Hitchcock, C. 2001. The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy* XCVIII(6):273–299.
- Hitchcock, C. 2007. Prevention, preemption, and the principle of sufficient reason. *Philosophical Review* 116:495–532.
- Hopkins, M., and Pearl, J. 2003. Clarifying the usage of structural models for commonsense causal reasoning. In *Proc. AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*.
- Hopkins, M. 2001. A proof of the conjunctive cause conjecture. Unpublished manuscript.
- Kahneman, D., and Miller, D. T. 1986. Norm theory: comparing reality to its alternatives. *Psychological Review* 94(2):136–153.
- Knobe, J., and Fraser, B. 2008. Causal judgment and moral judgment: two experiments. In Sinnott-Armstrong, W., ed., *Moral Psychology, Volume 2: The Cognitive Science of Morality*. Cambridge, MA: MIT Press. 441–447.
- Lagnado, D.; Gerstenberg, T.; and Zultan, R. 2013. Causal responsibility and counterfactuals. *Cognitive Science* 37:1036–1073.
- Livengood, J. 2013. Actual causation in simple voting scenarios. *Nous* 47(2):316–345.
- Spohn, W. 2008. Personal email.
- Weslake, B. 2014. A partial theory of actual causation. *The British Journal for the Philosophy of Science*. To appear.