Graded Causation and Defaults*

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

Christopher Hitchcock
California Institute of Technology
cricky@caltech.edu

August 2012

Abstract

This paper extends the account of actual causation offered by Halpern and Pearl [2005]. We show that this account yields the wrong judgment in certain classes of cases. We offer a revised definition that incorporates consideration of defaults, typicality, and normality. The revised definition takes actual causation to be both graded and comparative. We then apply our definition to a number of cases.

1 Introduction

This paper extends a recent account of *actual causation* due to Halpern and Pearl [2005] (HP hereafter). By "actual causation" (also called "singular causation", "token causation", or just "causation" in the literature), we mean that relation that we judge to hold when we affirm a statement that one particular event caused another. These claims are typically made retrospectively, and expressed in the past tense. For example, the following claims all describe relations of actual causation:

- Paula's putting poison in the tea caused Victoria to die.
- A lightning strike caused the forest fire.
- The ignition of the rum aboard the ship caused Lloyd's of London to suffer a large financial loss.

Actual causation is the target of analysis, for example, in David Lewis's classic paper "Causation" [Lewis 1973] (although he calls it simply "causation"). This relation has been of interest in philosophy and the law in part because of its connection with issues of moral and legal responsibility (see for example, Moore [2009] for a detailed discussion of these connections).

*For helpful comments and discussion, we would like to thank Luke Glynn, Franz Huber, Josh Knobe, Laurie Paul, Jim Woodward, members of the McDonnell Causal Learning Collaborative, and audience members at the Association of Symbolic Logic meeting (Chicago 2010), the Society for Exact Philosophy (Kansas City 2010), the Formal Epistemology Festival (Toronto 2010), the Workshop on Actual Causation (Konstanz 2010), Ohio University, California Institute of Technology, Rutgers University, and the University of California at San Diego.

†Supported in part by NSF grants IIS-0812045 and IIS-0911036, AFOSR grants FA9550-08-1-0438 and FA9550-05-1-0055, and ARO grant W911NF-09-1-0281.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **AUG 2012**	2. REPORT TYPE	3. DATES COVERED **00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE
Graded Causation and Defaults

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Cornell University,Department of Computer Science,Ithaca,NY,14853

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This paper extends the account of actual causation offered by Halpern and Pearl [2005]. We show that this account yields the wrong judgment in certain classes of cases. We offer a revised definition that incorporates consideration of defaults, typicality, and normality. The revised definition takes actual causation to be both graded and comparative. We then apply our definition to a number of cases.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**38**	

Getting an adequate definition of actual causation has proven exceedingly difficult. There are literally hundreds of papers in fields as diverse as philosophy, law, economics, physics, and computer science proposing and criticizing definitions of actual causation. The definition due to HP uses *structural equations*. The intuition behind this definition, which goes back to Hume [1748], is that A is a cause of B if, had A not happened, B would not have happened. As is well known, this definition is too naive. To take an example due to Wright [1985], suppose that Victoria, the victim, drinks a cup of tea poisoned by Paula, but before the poison takes effect, Sharon shoots Victoria, and she dies. We would like to call Sharon's shot the cause of the Victoria's death, but if Sharon hadn't shot, Victoria would have died in any case. HP deal with this by, roughly speaking, considering the contingency where Sharon does not shoot. Under that contingency, Victoria dies if Paula administers the poison, and otherwise does not. To prevent the poisoning from also being a cause of Paula's death, HP put some constraints on the contingencies that could be considered. There are a number of closely related definitions in the literature, including an earlier version by Halpern and Pearl [2001], as well as offerings from Pearl [2000], Hitchcock [2001], Woodward [2003], Hall [2007], and Glymour and Wimberly [2007]. With the exception of [Halpern and Pearl 2001], we do not discuss any of these alternative theories explicitly. Nonetheless, we expect that it is possible to extend these theories along much the same lines we propose for the HP theory.

The HP theory faces several problems. In particular, Hall [2007] and Hiddleston [2005] argue that the HP definition gives inappropriate answers in "bogus prevention" and "short circuit" cases, which seem to have structural equations isomorphic to ones where the HP definition gives the appropriate answer. This suggests that there must be more to actual causation than just the structural equations. One approach to solving these problems, suggested by Hall [2007], Halpern [2008], and Hitchcock [2007], is to incorporate considerations about about *defaults*, *typicality*, and *normality*. In the present paper, we develop this approach in greater detail.

Our approach lets us do much more than just deal with bogus prevention and short circuits.[1] It also allows us to move away from causality being an all or nothing assignment—either A is a cause of B or it is not—to a more "graded" notion of causality. We can then talk about one event being viewed as more of a cause than another. To the extent that we tend to view one event as "the" cause, it is because it is the one that is the "best" cause.

Our approach also allows us to make sense of cases where people disagree about claims of actual causation, despite being in agreement about the underlying structure of a particular case. Consider, for example, cases of *causation by omission*. Suppose that while a homeowner is on vacation, the weather is hot and dry, and the flowers in her garden die. Had her next door neighbor watered the flowers, they would not have died. Is the neighbor's failure to water the flowers an actual cause of the flowers' death? The structure of the case is clear: the flowers would not have died if the weather had not been so hot and dry, or if someone had watered them. Nonetheless, different writers disagree about whether this is a case of actual causation (see Section 7.2 for details and references). If these diverse authors are not disagreeing about the underlying structure of the case, what is the source of their disagreement about the actual causation claim? By building norms, typicality, and defaults into our account of actual causation, we can accommodate and explain these differing judgments.

The rest of the paper is organized as follows. In the next two sections, we review the causal modeling framework that we employ, and the HP definition of actual causation. More details, intuition, and

[1] Indeed, although dealing with bogus prevention was the motivation for this work, as we show in Section 7.1, we actually can deal with this problem by an appropriate choice of structural equations, without appealing to normality.

motivation can be found in [Halpern and Pearl 2005] and the references therein; the discussion here is largely taken from [Halpern 2008]. Readers who are already familiar with these may skim these sections. In Section 4, we briefly review some of the problems faced by the HP theory. In Section 5, we informally introduce the notions of *defaults*, *typicality*, and *normality*, and provide some further motivation for incorporating these notions into a theory of actual causation. Section 6 contains our formal treatment of these notions, and presents our revised, graded definition of actual causation. We conclude by applying the revised definition to a number of examples.

2 Causal Models

The HP approach models the world using random variables and their values. For example, if we are trying to determine whether a forest fire was caused by lightning or an arsonist, we can construct a model using three random variables:

- FF for forest fire, where $FF = 1$ if there is a forest fire and $FF = 0$ otherwise;
- L for lightning, where $L = 1$ if lightning occurred and $L = 0$ otherwise;
- M for match (dropped by arsonist), where $M = 1$ if the arsonist drops a lit match, and $M = 0$ otherwise.

The choice of random variables determines the language used to frame the situation. Although there is no "right" choice, clearly some choices are more appropriate than others. For example, when trying to determine the cause of the forest fire, if there is no random variable corresponding to the lightning in a model then, in that model, we cannot hope to conclude that lightning is a cause of the forest fire.

Some random variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. For example, to model the fact that if a match is lit or lightning strikes then a fire starts, we could use the random variables M, FF, and L as above, with the equation

$$FF = \max(L, M).$$

Since the value of FF is the maximum of the values of L and M, FF is 1 if either of L and M is 1. Alternatively, if a fire requires both a lightning strike *and* a dropped match (perhaps the wood is so wet that it needs two sources of fire to get going), the appropriate equation for FF would be

$$FF = \min(L, M);$$

the value of FF is the minimum of the values of L and M. The only way that $FF = 1$ is if both $L = 1$ and $M = 1$. For future reference, we call the model that uses the first equation the *disjunctive* model, and the one that uses the second equation the *conjunctive* model.

The equality signs in these equations should be thought of more like assignment statements in programming languages than normal algebraic equalities. For example, the first equation tells us that once we set the values of M and L, then the value of FF is set to their maximum. This relation is *not* symmetric; if a forest fire starts some other way, that does not force the value of either M or L to be 1. This asymmetry corresponds to the asymmetry in what Lewis [1979] calls "non-backtracking" counterfactuals. Suppose that there actually was no lightning, and the arsonist did not drop a match. Then (using non-backtracking counterfactuals), we would say that *if* lightning had struck or the arsonist had lit her match, then there would have been a fire. However, we would *not* say that if there had been a fire, then either lightning would have struck, or the arsonist would have lit her match.

These models are somewhat simplistic. Lightning does not always result in a fire, nor does dropping a lit match. One way of dealing with this would be to make the assignment statements probabilistic. For example, we could say that the probability that $FF = 1$ conditional on $L = 1$ is .8. This approach would lead to rather complicated definitions. It is much simpler to think of all the equations as being deterministic and, intuitively, use enough variables to capture all the conditions that determine whether there is a forest fire are captured by random variables. One way to do this is simply to add those variables explicitly. For example, we could add variables that talk about the dryness of the wood, the amount of undergrowth, the presence of sufficient oxygen, and so on. If a modeler does not want to add all these variables explicitly (the details may simply not be relevant to the analysis), another alternative is to use a single variable, say U, which intuitively incorporates all the relevant factors, without describing them explicitly. The value of U would determine whether the lightning strikes and whether the match is dropped by the arsonist. In this way of modeling things, U would take on four possible values of the form (i, j), where i and j are both either 0 or 1. Intuitively, i describes whether the external conditions are such that the lightning strikes (and encapsulates all the conditions, such as humidity and temperature, that affect whether the lightning strikes); and j describes whether the arsonist drops the match (and thus encapsulates the psychological conditions that determine whether the arsonist drops the match).

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. In the forest-fire example, the variables M, L, and FF are endogenous. However, we want to take as given that there is enough oxygen for the fire and that the wood is sufficiently dry to burn. In addition, we do not want to concern ourselves with the factors that make the arsonist drop the match or the factors that cause lightning. Thus we do not include endogenous variables for these factors. However, their effects are nonetheless incorporated into the exogenous variable(s).

Formally, a *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ 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 $\mathcal{S}$ 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*). As suggested above, in the forest-fire example, we have $\mathcal{U} = \{U\}$, where U is the exogenous variable, $\mathcal{R}(U)$ consists of the four possible values of U discussed earlier, $\mathcal{V} = \{FF, L, M\}$, and $\mathcal{R}(FF) = \mathcal{R}(L) = \mathcal{R}(M) = \{0, 1\}$.

$\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X such that

$$F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \to \mathcal{R}(X).$$

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 $X = U + Y$.[2] Thus, if $Y = 3$

[2] Again, the 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.

and $U = 2$, then $X = 5$, regardless of how Z is set.

In the running forest-fire example, where U has four possible values of the form (i, j), the i value determines the value of L and the j value determines the value of M. Although F_L gets as arguments the values of U, M, and FF, in fact, it depends only on the (first component of) the value of U; that is, $F_L((i, j), m, f) = i$. Similarly, $F_M((i, j), l, f) = j$. In this model, the value of FF depends only on the value of L and M. *How* it depends on them depends on whether we are considering the conjunctive model or the disjunctive model.

It is sometimes helpful to represent a causal model graphically. Each node in the graph corresponds to one variable in the model. An arrow from one node, say L, to another, say FF, indicates that the former variable figures as a nontrivial argument in the equation for the latter. Thus, we could represent either the conjunctive or the disjunctive model using Figure 2(a). Often we omit the exogenous variables from the graph; in this case, we would represent either model using Figure 2(b). Note that the graph conveys only the qualitative pattern of dependence; it does not tell us how one variable depends on others. Thus the graph alone does not allow us to distinguish between the disjunctive and the conjunctive models.

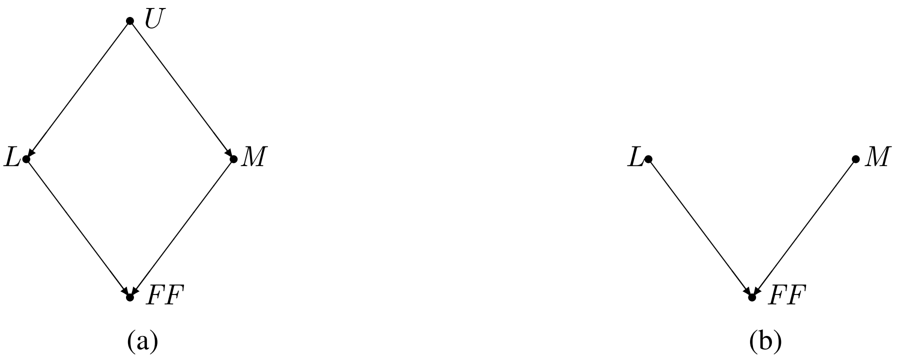

Figure 1: A graphical representation of structural equations.

The key role of the structural equations is to define what happens in the presence of external *interventions*. For example, we can explain what would happen if one were to intervene to prevent the arsonist from dropping the match. In the disjunctive model, there is a forest fire exactly exactly if there is lightning; in the conjunctive model, there is definitely no fire. Setting the value of some variable X to x in a causal model $M = (\mathcal{S}, \mathcal{F})$ by means of an intervention results in a new causal model denoted $M_{X=x}$. $M_{X=x}$ is identical to M, except that the equation for X in $\mathcal{F}$ is replaced by $X = x$. We sometimes talk of "fixing" the value of X at x, or "setting" the value of X to x. These expressions should also be understood as referring to interventions on the value of X. Note that an "intervention" does not necessarily imply human agency. The idea is rather that some independent process overrides the existing causal structure to determine the value of one or more variables, regardless of the value of its (or their) usual causes. Woodward [2003] gives a detailed account of such interventions. Lewis [1979] suggests that we think of the antecedents of non-backtracking counterfactuals as being made true by "small miracles". These "small miracles" would also count as interventions.

It may seem circular to use causal models, which clearly already encode causal information, to define actual causation. Nevertheless, there is no circularity. The models do not directly represent relations of *actual causation*. Rather, they encode information about what would happen under various possible interventions. Equivalently, they encode information about which non-backtracking counterfactuals are true. We will say that the causal represent (perhaps imperfectly or incompletely) the underlying *causal structure*. While there may be some freedom of choice regarding which variables are included in a model, once an appropriate set of variables has been selected, there should be an objective fact about which equations among those variables correctly describe the effects of interventions on some particular system of interest.[3] Perhaps it will be possible to analyze these relations completely in non-causal terms, as Lewis [1979] hoped. Perhaps it will not be. In either event, causal structure is distinct from actual causation, and an account of actual causation in terms of causal models is no more circular than Lewis's analysis of "causation" (really actual causation) in terms of non-backtracking counterfactuals.

In a causal model, it is possible that the value of X can depend on the value of Y (that is, the equation F_X is such that changes in Y can change the value of X) and the value of Y can depend on the value of X. Intuitively, this says that X can potentially affect Y and that Y can potentially affect X. While allowed by the framework, this type of situation does not happen in the examples of interest; dealing with it would complicate the exposition. Thus, for ease of exposition, we restrict attention here to what are called *recursive* (or *acyclic*) models. This is the special case where there is some total ordering $<$ of the endogenous variables (the ones in $\mathcal{V}$) such that if $X < Y$, then X is independent of Y, that is, $F_X(\ldots, y, \ldots) = F_X(\ldots, y', \ldots)$ for all $y, y' \in \mathcal{R}(Y)$. If $X < Y$, then the value of X may affect the value of Y, but the value of Y cannot affect the value of X.[4] Intuitively, if a theory is recursive, there is no feedback. The graph representing an acyclic causal model does not contain any directed paths leading from a variable back into itself, where a *directed path* is a sequence of arrows aligned tip to tail.

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 $<$. 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 $Y < X$, depend only on the exogenous variables, so their value is immediately determined by the values of the exogenous variables. The values of values later in the order can be determined once we have determined the values of all the variables earlier in the order.

There are many nontrivial decisions to be made when choosing the structural model to describe a given situation. One significant decision is the set of variables used. As we shall see, the events that can be causes and those that can be caused are expressed in terms of these variables, as are all the intermediate events. The choice of variables essentially determines the "language" of the discussion; new events cannot be created on the fly, so to speak. In our running example, the fact that there is no variable for unattended campfires means that the model does not allow us to consider unattended campfires as a cause of the forest fire.

Once the set of variables is chosen, the next step is to decide which are exogenous and which are endogenous. As we said earlier, the exogenous variables to some extent encode the background situation that we want to take for granted. Other implicit background assumptions are encoded in the structural equations themselves. Suppose that we are trying to decide whether a lightning bolt or a match was the cause of the forest fire, and we want to take for granted that there is sufficient oxygen in the air and the wood is dry. We could model the dryness of the wood by an exogenous variable D with values 0 (the

[3] Although as Halpern and Hitchcock [2010] note, some choices of variables do not give rise to well-defined equations. This would count against using that set of variables to model the system of interest.

[4] This standard restriction is also made by HP except in the appendix of [Halpern and Pearl 2005].

wood is wet) and 1 (the wood is dry).[5] By making D exogenous, its value is assumed to be given and out of the control of the modeler.[6] We could also take the amount of oxygen as an exogenous variable (for example, there could be a variable O with two values—0, for insufficient oxygen, and 1, for sufficient oxygen). Alternatively, we could choose not to model moisture and oxygen explicitly at all. By using the equation $FF = \max(L, M)$, we are saying that the wood will burn if the match is lit or lightning strikes. Thus, the equation is implicitly modeling our assumption that there is sufficient oxygen for the wood to burn.

It is not always straightforward to decide what the "right" causal model is in a given situation. In particular, there may be disagreement about which variables to use, and which should be exogenous and endogenous (although if the variables are well-chosen, it should at least be clear what the equations among them should be, at least if the behavior of the system is understood). These decisions often lie at the heart of determining actual causation in the real world. While the formalism presented here does not provide techniques to settle disputes about which causal model is the right one, at least it provides tools for carefully describing the differences between causal models, so that it should lead to more informed and principled decisions about those choices. (Again, see [Halpern and Hitchcock 2010] for further discussion of these issues.)

3 The HP Definition of Actual Causation

To make the definition of actual causality precise, it is helpful to have a formal language for counterfactuals and interventions. Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, 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 $\mathcal{S}$)* is one of the form $[Y_1 = y_1, \ldots, Y_k = y_k]\varphi$, where

- φ is a Boolean combination of primitive events,
- $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{V}$, and
- $y_i \in \mathcal{R}(Y_i)$.

Such a formula is abbreviated $[\vec{Y} = \vec{y}]\varphi$. The special case where $k = 0$ is abbreviated as φ. Intuitively, $[Y_1 = y_1, \ldots, Y_k = y_k]\varphi$ says that φ would hold if each Y_i were set to y_i by an intervention, for $i = 1, \ldots, k$.

A causal formula φ is true or false in a causal model, given a context. We write $(M, \vec{u}) \models \varphi$ if the causal formula φ 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 solution to the equations in M in context $\vec{u}$ (that is, the unique vector of values for the endogenous 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, $(M, \vec{u}) \models [\vec{Y} = \vec{y}]\varphi$ if $(M_{\vec{Y}=\vec{y}}, \vec{u}) \models \varphi$.

For example, if M is the disjunctive causal model for the forest fire, and u is the context where there is lightning and the arsonist drops the lit match, then $(M, u) \models [M = 0](FF = 1)$, since even if

[5] Of course, in practice, we may want to allow D to have more values, indicating the degree of dryness of the wood, but that level of complexity is unnecessary for the points we are trying to make here.

[6] See Section 7.4 for an alternative way of dealing with events that we want to take as given.

the arsonist is somehow prevented from dropping the match, the forest burns (thanks to the lightning); similarly, $(M, u) \models [L = 0](FF = 1)$. However, $(M, u) \models [L = 0; M = 0](FF = 0)$: if the arsonist does not drop the lit match and the lightning does not strike, then the forest does not burn.

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. This idea goes back to at least Hume [1748, Section VIII], who said:

> We may define a cause to be an object followed by another, ..., if the first object had not been, the second never had existed.

This is essentially the *but-for* test, perhaps the most widely used test of actual causation in tort adjudication. The but-for test states that an act is a cause of injury if and only if, but for the act (i.e., had the the act not occurred), the injury would not have occurred. David Lewis [1973] has also advocated a counterfactual definition of causation.

There are two well-known problems with this definition. The first can be seen by considering the disjunctive causal model for the forest fire again. Suppose that the arsonist drops a match and lightning strikes. Which is the cause? According to a naive interpretation of the counterfactual definition, neither is. If the match hadn't dropped, then the lightning would still have struck, so there would have been a forest fire anyway. Similarly, if the lightning had not occurred, there still would have been a forest fire. As we shall see, the HP definition declares both lightning and the arsonist actual causes of the fire. (In general, there may be more than one actual cause of an outcome.)

A more subtle problem is that of *preemption*, where there are two potential causes of an event, one of which preempts the other. Preemption is illustrated by the following story, due to Hitchcock [2007]:

> An assassin puts poison in a victim's drink. If he hadn't poisoned the drink, a backup assassin would have. The victim drinks the poison and dies.

Common sense suggests that the assassin's poisoning the drink caused the victim to die. However, it does not satisfy the naive counterfactual definition either; if the assassin hadn't poisoned the drink, the backup would have, and the victim would have died anyway.

The HP definition deals with these problems by defining actual causation as counterfactual dependence *under certain contingencies*. In the forest-fire example, the forest fire *does* counterfactually depend on the lightning under the contingency that the arsonist does not drop the match; similarly, the forest fire depends counterfactually on the arsonist's match under the contingency that the lightning does not strike. In the poisoning example, the victim's death *does* dounterfactually depend on the first assassin's poisoning the drink under the contingency that the backup does not poison the drink (perhaps because she is not present). However, we need to be careful here to limit the contingencies that can be considered. We do not want to count the backup assassin's presence as an actual cause of death by considering the contingency where the first assassin does not poison the drink. We consider the first assassin's action to be the cause of death because it was her poison that the victim consumed. Somehow the definition must capture this obvious intuition. A big part of the challenge of providing an adequate definition of actual causation comes from trying get these restrictions just right.

With this background, we now give the HP definition of causality.[7] The definition is relative to a causal model (and a context); A may be a cause of B in one causal model but not in another. The definition consists of three clauses. The first and third are quite simple; all the work is going on in the second clause.

The types of events that the HP definition allows as actual causes are ones of the form $X_1 = x_1 \land \ldots \land X_k = x_k$—that is, conjunctions of primitive events; this is often abbreviated as $\vec{X} = \vec{x}$. The events that can be caused are arbitrary Boolean combinations of primitive events. The definition does not allow statements of the form "A or A' is a cause of B", although this could be treated as being equivalent to "either A is a cause of B or A' is a cause of B". On the other hand, statements such as "A is a cause of B or B'" are allowed; as we shall see, this is not equivalent to "either A is a cause of B or A is a cause of B'".

Definition 3.1: (Actual cause) [Halpern and Pearl 2005] $\vec{X} = \vec{x}$ 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 $\vec{W}$ with $\vec{X} \subseteq \vec{Z}$, and settings $\vec{x}'$ and $\vec{w}$ of the variables in $\vec{X}$ and $\vec{W}$, respectively, such that if $(M, \vec{u}) \models Z = z^*$ for all $Z \in \vec{Z}$, then both of the following conditions hold:

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

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

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

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 for changing φ in AC2(a) are considered part of a cause; inessential elements are pruned. Without AC3, if dropping a lit match qualified as a cause of the forest fire, then dropping a match and sneezing would also pass the tests of AC1 and AC2. AC3 serves here to strip "sneezing" and other irrelevant, over-specific details from the cause. Clearly, all the "action" in the definition occurs in AC2. 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(s) 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 truth 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; it may be necessary to intervene on the value of one or more variables in $\vec{W}$ to allow this effect to manifest itself. For example, consider the context

[7] In fact, this is actually a preliminary definition in [Halpern and Pearl 2005], although it is very close the final definition. We will ignore here the final modification, which will be supplanted by our new account. When we talk of "the HP definition", we should be understood as referring to Definition 3.1 below, rather than to the final definition in [Halpern and Pearl 2005].

where both the lightning strikes and the arsonist drops a match in the disjunctive model of the forest fire. Stopping the arsonist from dropping the match will not prevent the forest fire. The counterfactual effect of the arsonist on the forest fire manifests itself only in a situation where the lightning does not strike (i.e., where L is set to 0). AC2(a) is what allows us to call both the lightning and the arsonist causes of the forest fire.

AC2(b) is perhaps the most complicated condition. It limits the "permissiveness" of AC2(a) with regard to the contingencies that can be considered. Essentially, it ensures that the change in the value $\vec{X}$ 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 intervention on $\vec{W}$, this perturbation has no impact on the value of φ. We capture the fact that the perturbation has no impact on the value of φ by saying that if some variables Z on the causal path were set to their original values in the context $\vec{u}$, φ would still be true, as long as $\vec{X} = \vec{x}$. Note that it follows from AC2(b) that an intervention that sets the value of the variables in $\vec{W}$ to their *actual* values is always permissible. Such an intervention might still constitute a change to the model, since the value of one or more variables in $\vec{W}$ might otherwise change when we change the value of $\vec{X}$ from $\vec{x}$ to $\vec{x}'$. Note also that if φ counterfactually depends on $\vec{X} = \vec{x}$, AC rules that $\vec{X} = \vec{x}$ is an actual cause of φ (assuming that AC1 and AC3 are both satisfied). We can simply take $\vec{W} = \emptyset$; both clauses of AC2 are satisfied.

In an earlier paper, Halpern and Pearl [2001] offered a more permissive version of clause AC2(b). That clause required that $(M, \vec{u}) \models [\vec{X} = \vec{x}, \vec{W} = \vec{w}, \vec{Z}' = \vec{z}^*]\varphi$ for all subsets $\vec{Z}'$ of $\vec{Z}$. In other words, it required only that this relation hold for the specific setting $\vec{W} = \vec{w}$, and not that it hold for every setting $\vec{W}' = \vec{w}$ where $\vec{W}'$ is a subset of $\vec{W}$. The change was prompted by a counterexample due to Hopkins and Pearl [2003]. It turns out that this counterexample can also be dealt with in the same way that we deal with bogus prevention cases below (see Section 7.1). For continuity, however, we continue to use the more recent definition from [Halpern and Pearl 2005]. We briefly discuss some of the consequences of the alternate definition in the appendix.

We now show how the HP definition handles our two problematic cases.

Example 3.2: For the forest-fire example, let M be the disjunctive causal model for the forest fire sketched earlier, with endogenous variables L, M, and FF, and equation $FF = \max(L, M)$. Clearly $(M, (1,1)) \models FF = 1$ and $(M, (1,1)) \models L = 1$; in the context (1,1), the lightning strikes and the forest burns down. Thus, AC1 is satisfied. AC3 is trivially satisfied, since $\vec{X}$ consists of only one element, L, so must be minimal. For AC2, take $\vec{Z} = \{L, FF\}$ and take $\vec{W} = \{M\}$, let $x' = 0$, and let $w = 0$. Clearly, $(M, (1,1)) \models [L = 0, M = 0](FF \neq 1)$; if the lightning does not strike and the match is not dropped, the forest does not burn down, so AC2(a) is satisfied. To see the effect of the lightning, we must consider the contingency where the match is not dropped; the definition allows us to do that by setting M to 0. (Note that here setting L and M to 0 overrides the effects of U; this is critical.) Moreover, $(M, (1,1)) \models [L = 1, M = 0](FF = 1)$; if the lightning strikes, then the forest burns down even if the lit match is not dropped, so AC2(b) is satisfied. (Note that since $\vec{Z} = \{L, FF\}$, the only subsets of $\vec{Z} - \vec{X}$ are the empty set and the singleton set consisting of just FF.) As this example shows, the HP definition need not pick out a unique actual cause; there may be more than one actual cause of a given outcome. ▮

Example 3.3: For the poisoning example, we can include in our causal model M the following endogenous variables:

- A = 1 if the assassin poisons the drink, 0 if not;
- R = 1 if the backup is ready to poison the drink if necessary, 0 if not;
- B = 1 if the backup poisons the drink, 0 if not;
- D = 1 if the victim dies, 0 if not.

We also have an exogenous variable U that determines whether the first assassin poisons the drink and whether the second assassin is ready. Let U have four values of the form (u_1, u_2) with $u_i \in \{0, 1\}$ for $i = 1, 2$. The equations are

$$\begin{aligned} &A = u_1; \\ &R = u_2; \\ &B = (1 - A) \times R; \\ &D = \max(A, B). \end{aligned}$$

The third equation says that the backup poisons the drink if she is ready and the first assassin doesn't poison the drink. The fourth equation says that the victim dies if either assassin poisons the drink. This model is represented graphically in Figure 3.3. In the actual context, where $U = (1, 1)$, we have $A = 1, R = 1, B = 0$, and $D = 1$. We want our account to give the result that $A = 1$ is an actual cause of $D = 1$, while $R = 1$ is not.

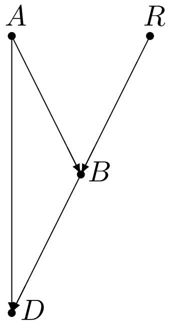

Figure 2: The poisoning example.

For the former, note that $D = 1$ does not counterfactually depend on $A = 1$: if A had been 0, B would have been 1, and D would still have been 1. Nonetheless, Definition AC rules that $A = 1$ is an actual cause of $D = 1$. It is easy to see that AC1 and AC3 are satisfied. $A = 1$ and $D = 1$ are both true in the actual context where $U = (1, 1)$, so AC1 is satisfied. Moreover, $A = 1$ is minimal, so AC3 is satisfied. For AC2, let $\vec{Z} = \{A, D\}$ and $\vec{W} = \{R, B\}$, with $\vec{x}' = 0$ and $\vec{w} = (1, 0)$ (i.e., $R = 1$ and $B = 0$). Checking AC2(a), we see that $(M, (1, 1)) \models [A = 0, R = 1, B = 0](D \neq 1)$. That is, when we intervene to set A to 0, R to 1, and B to 0, D takes a different value. While the victim's death does not counterfactually depend on the assassin's poisoning the drink, counterfactual dependence is restored if we also hold fixed that the backup did not act. Moreover, AC2(b) is also satisfied, for $\vec{W} = \vec{w}$ is just the setting $R = 1$ and $B = 0$, the values of R and B in the actual context. So if A also takes on its actual vaue of 1, then $D = 1$. Note that we could also choose $R = 0$ and $B = 0$ for the setting $\vec{W} = \vec{w}$. That is, it makes no difference to either part of AC2 if we intervene to prevent the backup from being ready. Alternately, we could include R in $\vec{Z}$ instead of $\vec{W}$; again the analysis is unaffected.

We now show that $R = 1$ is not an actual cause of $D = 1$. AC1 and AC3 are satisfied. In order to satisfy AC2(a), we need $\vec{x}' = 0, \vec{Z} = \{R, B, D\}, \vec{W} = \{A\}$, and $\vec{w}' = 0$. In words, the only way to get the victim's death to counterfactually depend on the backup's readiness is if we intervene to prevent the first assassin from poisoning the drink. But now we can verify that these settings do not also satisfy AC2(b). Since the actual value of B was 0, AC2(b) requires that in order for $A = 0$ to be an admissible setting of $\vec{W} = \{A\}$, we must have $(M, (1, 1)) \models [A = 0, B = 0]D = 1$. That is, in order for $A = 0$ to be an admissible setting, this setting must not change the value of D, even if variables such as B that are on the causal path from R to D are held fixed at their actual value. But this condition fails: $(M, (1, 1)) \models [A = 0, B = 0]D = 0$. Holding fixed that the backup did not poison the drink, if the assassin hadn't poisoned the drink either, the victim would have not have died. Intuitively, the idea is that the backup's readiness can be an actual cause of death only if the backup actually put poison in the drink. In this way, clause AC2(b) builds in the idea that there must be an appropriate sort of causal chain or process in order for $R = 1$ to be an actual cause of $D = 1$. This example also shows the importance of restricting the permissible contingencies that we can look at when re-evaluating counterfactual dependence. ∎

4 Problems for the HP Theory

While the definition of actual causation given in Definition 3.1 works well in many cases, it does not always deliver answers that agree with (most people's) intuition. Consider the following example of "bogus prevention", taken from Hitchcock [2007], based on an example due to Hiddleston [2005].

Example 4.1: Assassin is in possession of a lethal poison, but has a last-minute change of heart and refrains from putting it in Victim's coffee. Bodyguard puts antidote in the coffee, which would have neutralized the poison had there been any. Victim drinks the coffee and survives. Is Bodyguard's putting in the antidote a cause of Victim surviving? Most people would say no, but according to the HP definition, it is. For in the contingency where Assassin puts in the poison, Victim survives iff Bodyguard puts in the antidote. ∎

Example 4.1 illustrates what seems to be a significant problem with Definition 3.1. The structural equations for Example 4.1 are *isomorphic* to those in the disjunctive version of the forest-fire example, provided that we interpret the variables appropriately. Specifically, take the endogenous variables in Example 4.1 to be A (for "assassin does not put in poison"), B (for "bodyguard puts in antidote"), and VS (for "victim survives"). Then A, B, and VS satisfy exactly the same equations as L, M, and FF, respectively. In the context where there is lightning and the arsonists drops a lit match, both the the lightning and the match are causes of the forest fire, which seems reasonable. But here it does not seem reasonable that Bodyguard's putting in the antidote is a cause. Nevertheless, any definition that just depends on the structural equations is bound to give the same answers in these two examples.

A second type of case illustrating the same problem involves what Hall [2007] calls "short circuits". Hitchcock [2007] gives the following example (which he calls "CH" or "counterexample to Hitchcock", since it is a counterexample to the theory of [Hitchcock 2001]):

Example 4.2: A victim has two bodyguards charged with protecting him from assassination. The bodyguards conspire to make it appear as though they have foiled an attempted poisoning. They plan to put poison in victim's drink, and also to put in an antidote that will neutralize the poison. However, they do not want to put the poison in until they are certain that the antidote has safely been put in the drink. Thus, the first bodyguard adds antidote to the drink, and the second waits until the antidote has been added before adding the poison. If the first bodyguard were interrupted, or somehow prevented from putting the antidote in, the second would not add the poison. As it happens, both the antidote and the poison are added, so the poison is neutralized; the victim drinks the harmless liquid and lives. ∎

Most people, although by no means all, judge that putting the antidote into the drink is not an actual cause of the victim's survival. Put another way, administering the antidote did not prevent death. This is because putting the antidote in the drink caused the very threat that it was meant to negate. If the antidote weren't put in, there would have been no poison to neutralize. However, it turns out that this example has a structure that is isomorphic to the preemption example discussed earlier (Example 3.3). This is not immediately obvious; we discuss the technical details in Section 7.7. For now, it suffices to note that if we hold fixed that the second bodyguard puts in the poison, then intervening on whether the first bodyguard puts in the antidote makes a difference for whether the victim dies.

In both kinds of case, examples that have isomorphic structural equation models are judged to have different relations of actual causation. This suggests that there must be more to actual causation than just the structural equations.

5 Defaults, Typicality, and Normality

Our revised account of actual causation incorporates the concepts of *defaults*, *typicality*, and *normality*. These are related, although somewhat different notions:

- A *default* is an assumption about what happens, or what is the case, when no additional information is given. For example, we might have as a default assumption that birds fly. If we are told that Tweety is a bird, and given no further information about Tweety, then it is natural to infer that Tweety flies. Such inferences are *defeasible*: they can be overridden by further information. If we are additionally told that Tweety is a penguin, we retract our conclusion that Tweety flies. Default logics (see, e.g., [Marek and Truszczyński 1993; Reiter 1980; Reiter 1987]) attempt to capture the structure of these kinds of inferences.

- To say that birds *typically* fly is to say not merely that flight is statistically prevalent among birds, but also that flight is characteristic of the type "bird". Even though not all birds do fly, flying is something that we do characteristically associate with birds.

- The word *normal* is interestingly ambiguous. It seems to have both a descriptive and a prescriptive dimension. To say that something is normal in the descriptive sense is to say that it is the statistical mode or mean (or close to it). On the other hand, we often use the shorter form *norm* in a more prescriptive sense. To conform with a norm is to follow a prescriptive rule. Prescriptive norms can take many forms. Some norms are moral: to violate them would be to perform a moral wrong. For example, many people believe that there are situations in which it would be wrong to lie, even if there are no laws or explicit rules forbidding this behavior. Laws are another kind of norm, adopted for the regulation of societies. Policies that are adopted by institutions can also be norms. For instance, a company may have a policy that employees are not allowed to be absent from

> work unless they have a note from their doctor. There can also be norms of proper functioning in machines or organisms. There are specific ways that human hearts and car engines are *supposed* to work, where "supposed" here has not merely an epistemic force, but a kind of normative force. Of course, a car engine that does not work properly is not guilty of a moral wrong, but there is nonetheless a sense in which it fails to live up to a certain kind of standard.

While this might seem like a heterogeneous mix of concepts, they are intertwined in a number of ways. For example, default inferences are successful just to the extent that the default is normal in the statistical sense. Adopting the default assumption that a bird can fly facilitates successful inferences in part because most birds are able to fly. Similarly, we classify objects into types in part to group objects into classes most of whose members share certain features. Thus, the type "bird" is useful partly because there is a suite of characteristics shared by most birds, including the ability to fly. The relationship between the statistical and prescriptive senses of "normal" is more subtle. It is, of course, quite possible for a majority of individuals to act in violation of a given moral or legal norm. Nonetheless, we think that the different kinds of norm often serve as heuristic substitutes for one another. For example, well-known experiments by Kahneman and Tversky [Kahneman and Tversky 1982; Tversky and Kahneman 1973] show that we are often poor at explicit statistical reasoning, employing instead a variety of heuristics. Rather than tracking statistics about how many individuals behave in a certain way, we might well reason about how people ought to behave in certain situations. The idea is that we use a script or a template for reasoning about certain situations, rather than actual statistics. Prescriptive norms of various sorts can play a role in the construction of such scripts. It is true, of course, that conflation of the different sorts of norm can sometimes have harmful consequences. Less than a hundred years ago, for example, left-handed students were often forced to learn to write with their right hands. In retrospect, this looks like an obviously fallacious inference from the premise that the majority of people write with their right hand to the conclusion that it is somehow wrong to write with the left hand. But the very ease with which such an inference was made illustrates the extent to which we find it natural to glide between the different senses of "norm".

That there should be a connection between defaults, norms, typicality, and causality is not a new observation. Kahneman and Tversky [1982], Kahneman and Miller [1986], and others have shown that both statistical and prescriptive norms can affect counterfactual reasoning. For example, Kahneman and Miller [1986, p. 143] point out that "an event is more likely to be undone by altering exceptional than routine aspects of the causal chain that led to it." Given the close connection between counterfactual reasoning and causal reasoning, this suggests that norms will also affect causal judgment. Recent experiments have confirmed this. Alicke [1992] shows that subjects are more likely to judge that someone caused some negative outcome when they have a negative evaluation of that person. Cushman, Knobe, and Sinnott-Armstrong [2008] have shown that subjects are more likely to judge that an agent's action causes some outcome when they hold that the action is morally wrong; Knobe and Fraser [2008] have shown that subjects are more likely to judge that an action causes some outcome if it violates a policy; and Hitchcock and Knobe [2009] have shown that this effect occurs with norms of proper functioning.

Many readers are likely to be concerned that incorporating considerations of normality and typicality into an account of actual causation will have a number of unpalatable consequences. Causation is supposed to be an objective feature of the world. But while statistical norms are, arguably, objective, other kinds of norm do not seem to be. More specifically, the worry is that the incorporation of norms will render causation: (1) subjective, (2) socially constructed, (3) value-laden, (4) context-dependent, and (5) vague. It would make causation subjective because different people might disagree about what is typical or normal. For example, if moral values are not objective, then any effect of moral norms on causation will render causation subjective. Since some norms, such as laws, or policies implemented within institutions, are socially constructed, causation would become socially constructed too. Since moral norms, in particular, incorporate values, causation would become value-laden. Since there are many different kinds of norm, and they may sometimes come into conflict, conversational context will sometimes have to determine what is considered normal. This would render causation context-dependent. Finally, since typicality and normality seem to admit of degrees, this would render causation vague. But causation should be none of these things.

We believe that these worries are misplaced. While our account of *actual causation* incorporates all of these elements, actual causation is the wrong place to look for objectivity. *Causal structure*, as represented in the equations of a causal model, is objective. More precisely, once a suitable set of variables has been chosen,[8] there is an objectively correct set of structural equations among those variables. Actual causation, by contrast, is a fairly specialized causal notion. Actual causation involves the *post hoc* attribution of causal responsibility for some outcome. It is particularly relevant to making determinations of moral or legal responsibility. Hitchcock and Knobe [2009] argue that attributions of actual causation typically serve to identify appropriate targets of corrective intervention. Given the role it is supposed to play, it is nWot at all inappropriate for actual causation to have a subjective, normative dimension.

The philosophy literature tends to identify causation in general with actual causation. We believe that this identification is inappropriate. The confusion arises, in part, from the fact that in natural language we express judgments of actual causation using the simple verb "cause" (typically in the past tense). We say, for example, that the lightning caused the fire, the assassin's poisoning the drink caused the victim to die, and that one neuron's firing caused another to fire. Our language gives us no clue that it is a rather specialized causal notion that we are deploying.

While we accept that the assignment of norms and defaults is subjective, context-dependent, and so on, we do not think that it is completely unconstrained. In any given causal model, certain assignments of default values to variables are more natural, or better motivated, than others. An assignment of norms and defaults is the sort of thing that can be defended and criticized on rational grounds. For example, if a lawyer were to argue that a defendant did or did not cause some harm, she would have to give arguments in support of her assignments of norms or defaults. (See [Halpern and Hitchcock 2010] for further discussion of this issue.)

6 Extended Causal Models

Following Halpern [2008], we deal with the problems raised in Section 4 by assuming that an agent has, in addition to a theory of causal structure (as modeled by the structural equations), a theory of "normality" or "typicality". This theory would include statements like "typically, people do not put poison in coffee". There are many ways of giving semantics to such typicality statements, including *preferential structures* [Kraus, Lehmann, and Magidor 1990; Shoham 1987], *ϵ-semantics* [Adams 1975; Geffner 1992; Pearl 1989], *possibilistic structures* [Dubois and Prade 1991], and ranking functions [Goldszmidt and Pearl 1992; Spohn 1988]. Halpern [2008] used the last approach, but it builds in an assumption that the normality order on worlds is total. As we show by example, this does not allow us

[8] See [Halpern and Hitchcock 2010] for discussion of what makes for a suitable choice of variables

to deal with some examples. Thus, we use a slightly more general approach here, based on preferential structures.

Take a *world* to be an assignment of values to all the exogenous variables in a causal model.[9] Intuitively, a world is a complete description of a situation given the language determined by the set of endogenous variables. Thus, a world in the forest-fire example might be one where $M = 1$, $L = 0$, and $FF = 0$; the match is dropped, there is no lightning, and no forest fire. As this example shows, a "world" does not have to satisfy the equations of the causal model.

For ease of exposition, in the rest of the paper, we make a somewhat arbitrary stipulation regarding terminology. In what follows, we use "default" and "typical" when talking about individual variables or equations. For example, we might say that the default value of a variable is zero, or that one variable typically depends on another in a certain way. We use "normal" when talking about worlds. Thus, we say that one world is more normal than another. In the present paper, we do not develop a formal theory of typicality, but assume only that typical values for a variable are influenced by the kinds of factors discussed in the previous section. We also assume that it is typical for endogenous variables to be related to one another in the way described by the structural equations of a model, unless there is some specific reason to think otherwise. The point of this assumption is to ensure that the downstream consequences of what is typical are themselves typical (again, in the absence of any specific reason to think otherwise).

In contrast to our informal treatment of defaults and typicality, we provide a formal representation of normality. We assume that there is a partial preorder $\succeq$ over worlds; $s \succeq s'$ means that world s is at least as normal as world s'. The fact that $\succeq$ is a partial preorder means that it is reflexive (for all worlds s, we have $s \succeq s$, so s is at least as normal as itself) and transitive (if s is at least as normal as s' and s' is at least as normal as s'', then s is at least as normal as s'').[10] We write $s \succ s'$ if $s \succeq s'$ and it is not the case that $s' \succeq s$, and $s \equiv s'$ if $s \succeq s'$ and $s' \succeq s$. Thus, $s \succ s'$ means that s is strictly more normal than s', while $s \equiv s'$ means that s and s' are equally normal. Note that we are not assuming that $\succeq$ is total; it is quite possible that there are two worlds s and s' that are incomparable as far as normality. The fact that s and s' are incomparable does *not* mean that s and s' are equally normal. We can interpret it as saying that the agent is not prepared to declare either s or s' as more normal than the other, and also not prepared to say that they are equally normal; they simply cannot be compared in terms of normality.

One important issue concerns the relationship between *typicality* and *normality*. Ideally, one would like to have a sort of compositional semantics. That is, given a set of statements about the typical values of particular variables and a causal model, a normality ranking on worlds could be generated that in some sense respects those statements. We develop such an account in a companion paper [Halpern and Hitchcock 2012]. In the present paper, we make do with a few rough-and-ready guidelines. Suppose that s and s' are worlds, that there is some nonempty set $\vec{X}$ of variables that take more typical values in s than they do in s', and no variables that take more typical values in s' than in s; then $s \succ s'$. However, if there is both a nonempty $\vec{X}$ set of variables that take more typical values in s than they do in s', and a nonempty set $\vec{Y}$ of variables that take more typical values in s' than they do in s, then s and s' are

[9] It might be apt to use "small world" to describe such an assignment, to distinguish it from a "large world", which would be an assignment of values to all of the variables in a model, both endogenous and exogenous. While there may well be applications where large worlds are needed, the current application requires only small worlds. The reason for this is that all of the worlds that are relevant to assessing actual causation in a specific context $\vec{u}$ result from intervening on endogenous variables, while leaving the exogenous variables unchanged.

[10] If $\succeq$ were a partial order rather than just a partial preorder, it would satisfy an additional assumption, *antisymmetry*: $s \succeq s'$ and $s' \succeq s$ would have to imply $s = s'$. This is an assumption we do *not* want to make.

incomparable, unless there are special considerations that would allow us to rank them. This might be in the form of statement that $\vec{x}$ is a more a typical setting for $\vec{X}$ than $\vec{y}$ is of $\vec{Y}$. We consider an example where such a rule seems very natural in Section 7.6 below.

Take an *extended causal model* to be a tuple $M = (\mathcal{S}, \mathcal{F}, \succeq)$, where $(\mathcal{S}, \mathcal{F})$ is a causal model, and $\succeq$ is a partial preorder on worlds, which can be used to compare how normal different worlds are. In particular, $\succeq$ can be used to compare the actual world to a world where some interventions have been made. Which world is the actual world? That depends on the context. In an acyclic extended causal model, a context $\vec{u}$ determines a world denoted $s_{\vec{u}}$. We can think of the world $s_{\vec{u}}$ as the actual world, given context $\vec{u}$, since it is the world that would occur given the setting of the exogenous variables in $\vec{u}$, provided that there are no external interventions.

We can now modify Definition 3.1 slightly to take the ranking of worlds into account by taking $\vec{X} = \vec{x}$ to be a *cause of φ in an extended model M and context $\vec{u}$* if $\vec{X} = \vec{x}$ is a cause of φ according to Definition 3.1, except that in AC2(a), and $s_{\vec{X}=\vec{x}', \vec{W}=\vec{w}, \vec{u}} \succeq s_{\vec{u}}$, 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 $\vec{u}$. This can be viewed as a formalization of Kahneman and Miller's [1986] observation that we tend to consider only possibilities that result from altering atypical features of a world to make them more typical, rather than vice versa. In our formulation, worlds that result from interventions on the actual world "come into play" in AC2(a) only if they are at least as normal as the actual world.

This gives us a definition of causality that, as we shall see, lets us deal with a number of problems in the HP definition. But we can go further: we can use normality to rank actual causes. Doing so lets us explain the responses that people make to queries regarding actual causation. For example, although the HP approach allows for multiple causes of an outcome φ, people typically mention only one of them when asked for a cause. We would argue that they are picking the *best* cause, where best is judged in terms of normality. We now make this precise.

Say that world s is a *witness* for $\vec{X} = \vec{x}$ being a cause of φ in context $\vec{u}$ if for some choice of $\vec{Z}$, $\vec{W}$, $\vec{x}'$, and $\vec{w}$ for which AC2(a) and AC2(b) hold, s is the assignment of values to the endogenous variables that results from setting $\vec{X} = \vec{x}'$ and $\vec{W} = \vec{w}$ in context $\vec{u}$. In other worlds, a witness s is a world that demonstrates that AC2(a) holds. In general, there may be many witnesses for $\vec{X} = \vec{x}$ being a cause of φ. Say that s is a *best witness* for $\vec{X} = \vec{x}$ being a cause of φ if there is no other witness s' 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. We expect that someone's willingness to judge that $\vec{X} = \vec{x}$ is an actual cause of φ increases as a function of the normality of the best witness for $\vec{X} = \vec{x}$ in comparison to the best witness for other candidate causes. Thus, we are less inclined to judge that $\vec{X} = \vec{x}$ is an actual cause of φ when there are other candidate causes of equal or higher rank.

7 Examples

In this section, we give a number of examples of the power of this definition. (For simplicity, we omit explicit reference to the exogenous variables in the discussions that follow.) We start by considering the example that motivated this line of research.

7.1 Bogus prevention

Consider the bogus prevention problem of Example 4.1. Suppose that we use a causal model with three random variables:

- $A = 1$ if Assassin does puts in the poison, 0 if he does not;
- $B = 1$ if Bodyguard adds the antidote, 0 if he does not;
- $VS = 1$ if the victim survives, 0 if he does not.

Then the equation for VS is

$$VS = \max((1 - A), B).$$

A, B, and VS satisfy exactly the same equations as $1 - L$, M, and FF, respectively in Example 3.2. In the context where there is lightning and the arsonists drops a lit match, both the the lightning and the match are causes of the forest fire, which seems reasonable. Not surprisingly, the original HP definition declares both $A = 0$ and $B = 1$ to be actual causes of $VS = 1$. But here it does not seem reasonable that Bodyguard's putting in the antidote is a cause.

Using normality gives us a straightforward way of dealing with the problem. In the actual world, $A = 0, B = 1$, and $VS = 1$. The witness for $B = 1$ to be an actual cause of $VS = 1$ is the world where $A = 1, B = 0$, and $VS = 0$. If we make the assumption that both A and B typically take the value 0,[11] and make the assumptions about the relation between typicality and normality discussed in Section 6, this leads to a normality ordering in which the two worlds $(A = 0, B = 1, VS = 1)$ and $(A = 1, B = 0, VS = 0)$ are *incomparable*. Since the unique witness for $B = 1$ to be an actual cause of $VS = 1$ is incomparable with the actual world, our modified definition rules that $B = 1$ is not an actual cause of $VS = 1$. Interestingly, our account also rules that $A = 0$ is not an actual cause, since it has the same witness. This does not strike us as especially counterintuitive. (See the discussion of causation by omission in the following section.)

This example illustrates an advantage of the present account over the one offered in [Halpern 2008], in which normality is characterized by a total order. With a total order, we cannot declare $(A = 1, B = 0, VS = 0)$ and $(A = 0, B = 1, VS = 1)$ to be incomparable; we must compare them. To argue $A = 1$ is not a cause, we have to assume that $(A = 0, B = 1, VS = 1)$ is less normal than $(A = 1, B = 0, VS = 0)$. This ordering does not seem so natural.

As we hinted earlier, there is another, arguably preferable, way to handle this using the original HP definition, without appealing to normality. Suppose we add a variable PN to the model, representing whether a chemical reaction takes place in which poison is neutralized. The model has the following equations:

$$\begin{aligned} &PN = A \times B; \\ &VS = \max((1 - A), PN). \end{aligned}$$

This model is shown graphically in Figure 7.1. It is isomorphic to the model in Example 3.3, except that A and $1 - A$ are reversed. In this new model, $B = 1$ fails to be an actual cause of $VS = 1$ for the same reason the backup's readiness was not a cause of the victim's death in Example 3.3. By adding PN to

[11] Some readers have suggested that it would not be atypical for an assassin to poison the victim's drink. That is what assassins do, after all. Nonetheless, the action is morally wrong and unusual from the victim's perspective, both of which would tend to make it atypical.

the model, we can capture the intuition that the antidote doesn't count as a cause of survival unless it actually neutralized poison.

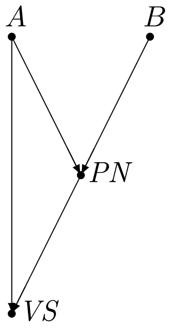

Figure 3: Another model of bogus prevention.

Despite the fact that we do not need normality for bogus prevention, it is useful in many other examples, as we show in the remainder of this section.

7.2 Omissions

As we mentioned in Section 5, there is disagreement in the literature over whether causation by omission should count as actual causation, despite the fact that there is no disagreement regarding the underlying causal structure. We can distinguish (at least) four different viewpoints in the flower-watering example described in the introduction:

(a) Beebee [2004] and Moore [2009], for example, argue against the existence of causation by omission in general;

(b) Lewis [2000, 2004] and Schaffer [2000, 2004] argue that omissions are genuine causes in such cases;

(c) Dowe [2000] and Hall [2004] argue that omissions have a kind of secondary causal status; and

(d) McGrath [2005] argues that the causal status of omissions depends on their normative status: whether the neighbor's omission caused the flowers to die depends on whether the neighbor was *supposed* to water the flowers.

Our account of actual causation can capture all four viewpoints. The obvious causal structure has three endogenous variables:

- $H = 1$ if the weather is hot, 0 if it is cool;
- $W = 1$ if the neighbor waters the flowers, 0 otherwise;
- $D = 1$ if the flowers die, 0 if they do not.

There is one equation:

$$D = H \times (1 - W).$$

The exogenous variables are such that $H = 1$ and $W = 0$, hence in the actual world $H = 1$, $W = 0$, and $D = 1$. The original HP definition rules that both $H = 1$ and $W = 0$ are actual causes of $D = 1$. The witness for $H = 1$ being a cause is the world $(H = 0, W = 0, D = 0)$, while the witness for $W = 0$ is $(H = 1, W = 1, D = 0)$. We claim that the difference between the viewpoints mentioned above can be understood as a disagreement either about the appropriate normality ranking, or the effect of graded causality.

Those who maintain that omissions are never causes can be understood as having a normality ranking where absences or omissions are always more typical than positive events. That is, the typical value for both H and W is 0. This ranking reflects a certain metaphysical view: there is a fundamental distinction between positive events and mere absences, and in the context of causal attribution, absences are always considered typical for candidate causes. This gives rise to a normality ranking where

$$(H = 0, W = 0, D = 0) \succ (H = 1, W = 0, D = 1) \succ (H = 1, W = 1, D = 0).$$

The fact that $(H = 0, W = 0, D = 0) \succ (H = 1, W = 0, D = 1)$ means that we can take $(H = 0, W = 0, D = 0)$ as a witness for $H = 1$ being a cause of $D = 1$. Indeed, most people would agree that the hot weather was a cause of the plants dying. Note that $(H = 1, W = 1, D = 0)$ is the witness for $W = 0$ being a cause of $D = 1$. If we take the actual world $(H = 1, W = 0, D = 1)$ to be more normal than this witness, intuitively, treating not acting as more normal than acting, then we cannot view $W = 0$ as an actual cause.

It should be clear that always treating "not acting" as more normal than acting leads to not allowing causation by omission. However, one potential problem for this sort of view is that it is not always clear what counts as a positive event, and what as a mere omission. For example, is holding one's breath a positive event, or is it merely the absence of breathing? If an action hero survives an encounter with poison gas by holding her breath, is this a case of (causation by) omission?

An advocate of the third viewpoint, that omissions have a kind of secondary causal status, may be interpreted as allowing a normality ordering of the form

$$(H = 0, W = 0, D = 0) \succ (H = 1, W = 1, D = 0) \succeq (H = 1, W = 0, D = 1).$$

This theory allows watering the plants to be as normal as not watering them, and hence $W = 0$ can be an actual cause of $D = 1$. However, $H = 1$ has a more normal witness, so under the comparative view, $H = 1$ is a much better cause than $W = 0$. On this view, then, we would be more strongly inclined to judge that $H = 1$ is an actual cause than that $W = 0$ is an actual cause. However, unlike those who deny that $W = 0$ is a cause of any kind, advocates of the third position might maintain that since $W = 0$ has a witness, it has some kind of causal status, albeit of a secondary kind.

An advocate of the second viewpoint, that omissions are always causes, could have essentially the same ordering as an advocate of the second viewpoint, but would take the gap between $(H = 0, W = 0, D = 0)$ and $(H = 1, W = 1, D = 0)$ to be much smaller, perhaps even allowing that $(H = 0, W = 0, D = 0) \equiv (H = 1, W = 1, D = 0)$. Indeed, if $(H = 0, W = 0, D = 0) \equiv (H = 1, W = 1, D = 0)$, then $H = 1$ and $W = 0$ are equally good candidates for being actual causes of $D = 1$. But note that with this viewpoint, not only is the neighbor who was asked to water the plants but did not a cause, so are all the other neighbors who were not asked. Moreover, the second or third viewpoints, if applied consistently to cases of bogus prevention, can not rule that bogus preventers are not actual causes of some sort. (This may give us an additional reason for treating bogus prevention as a type of preemption, as suggested in Section 7.1.)

The fourth viewpoint is that the causal status of omissions depends on their normative status. For example, suppose the neighbor had promised to water the homeowners' flowers; or suppose the two neighbors had a standing agreement to water each others' flowers while the other was away; or that the neighbor saw the wilting flowers, knew how much the homeowner loved her flowers, and could have watered them at very little cost or trouble to herself. In any of these cases, we might judge that the neighbor's failure to water the flowers was a cause of their death. On the other hand, if there was no reason to think that the neighbor had an obligation to water the flowers, or no reasonable expectation that she would have done so (perhaps because she, too, was out of town), then we would not count her omission as a cause of the flowers' death.

On this view, the existence of a norm, obligation, or expectation regarding the neighbor's behavior has an effect on whether the world $(H = 1, W = 1, D = 0)$ is considered at least as normal as the actual world $(H = 1, W = 0, D = 1)$. If there is no reason to expect the neighbor to water the flowers, then it is not normal. On the other hand, if the neighbor was supposed to water the flowers, then it will. This viewpoint allows us to explain why not all neighbors' failures to water the flowers should be treated equally.

7.3 Knobe effects

In a series of papers (e.g., [Cushman, Knobe, and Sinnott-Armstrong 2008; Hitchcock and Knobe 2009; Knobe and Fraser 2008]), Joshua Knobe and his collaborators have demonstrated that norms can influence our attributions of actual causation. For example, consider the following vignette, drawn from Knobe and Fraser [2008]:

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist repeatedly e-mailed them reminders that only administrators are allowed to take the pens. On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later, that day, the receptionist needs to take an important message ... but she has a problem. There are no pens left on her desk.

Subjects were then randomly presented with one of the following propositions, and asked to rank their agreement on a seven point scale from -3 (completely disagree) to +3 (completely agree):

> Professor Smith caused the problem.
> The administrative assistant caused the problem.

Subjects gave an average rating of 2.2 to the first claim, indicating agreement, but -1.2 to the second claim, indicating disagreement. Thus subjects are judging the two claims differently, due to the different normative status of the two actions. (Note that subjects were only presented with one of these claims: they were not given a forced choice between the two.)

If subjects are presented with a similar vignette, but where both groups are allowed to take pens, then subjects tend to give intermediate values. That is, when the vignette is changed so that Professor Smith is not violating a norm when he takes the pen, not only are subjects less inclined to judge that Professor Smith caused the problem, but they are more inclined to judge that the administrative assistant caused the problem.[12] This is interesting, since the status of the administrative assistant's action has not been changed. The most plausible interpretation of this result is that subjects' increased willingness to say that the administrative assistant caused the problem is a direct result of their decreased willingness to say that Professor Smith caused the problem. This suggests that attributions of actual causation are at least partly a comparative affair.

The obvious causal model of the original vignette has three random variables:

- $PT = 1$ if Professor S takes the pen, 0 if she does not;
- $AT = 1$ if the administrative assistant takes the pen, 0 if she does not;
- $PO = 1$ if the receptionist is unable to take a message, 0 if no problem occurs.

There is one equation:

$$PO = \min(PT, AT).$$

The exogenous variables are such that PT and AT are both 1. Therefore, in the actual world, we have $PT = 1$, $AT = 1$, and $PO = 1$.

The HP definition straightforwardly rules that $PT = 1$ and $AT = 1$ are both actual causes of $PO = 1$. (In both cases, we can let $\vec{W}$ be the empty set.) The best witness for $PT = 1$ being a cause is $(PT = 0, AT = 1, PO = 0)$; the best witness for $AT = 1$ being a cause is $(PT = 1, AT = 0, PO = 0)$. The original story suggests that the witness for $PT = 1$ being a cause is more normal than the witness for $AT = 1$, since administrative assistants are allowed to take pens, but professors are supposed to buy their own. So our account predicts that we are more strongly inclined to judge that $PT = 1$ is an actual cause. On the other hand, if the vignette does not specify that one of the actions violates a norm, we would expect the relative normality of the two witnesses to be much closer, which is reflected in how subjects actually rated the causes.

7.4 Causes vs. background conditions

It is common to distinguish between causes of some outcome, and mere background conditions that are necessary for that outcome (e.g. [Hart and Honoré 1985]). A standard example is a fire that is caused by a lit match. While the fire would not have occurred without the presence of oxygen in the atmosphere, the oxygen is deemed to be a background condition, rather than a cause. The HP definition allows us to make the distinction by taking the random variable representing the presence of oxygen to be exogenous; it then cannot be a cause. As we now show, making use of normality allows us to avoid having to make this exogenous/endogenous distinction.

We have three variables:

[12] Sytsma, Livengood, and Rose [2010] conducted the experiments. They had their subjects rate their agreement on 7-point scale from 1 (completely disagree) to 7 (completely agree). When they repeated Knobe and Fraser's original experiment, they got a rating of 4.05 for Professor Smith, and 2.51 for the administrative assistant. While their difference is less dramatic than Knobe and Fraser's, it is still statistically significant. When they altered the vignette so that Professor Smith's action was permissible, subjects gave an average rating of 3.0 for Professor Smith, and 3.53 for the administrative assistant.

- $M = 1$ if the match is lit, 0 if it is not lit;
- $O = 1$ if oxygen is present, 0 if there is no oxygen;
- $F = 1$ if there is a fire, 0 if there is no fire.

There is one equation:

$$F = \min(M, O).$$

The exogenous variables are such that in the actual world, $M = 1$ and $O = 1$, so $F = 1$.

Again, $F = 1$ counterfactually depends on both $M = 1$ and $O = 1$, so the HP definition rules that both are actual causes of $F = 1$. The witness for $M = 1$ being a cause is the world $(M = 0, O = 1, F = 0)$; the witness for $O = 1$ being a cause is $(M = 1, O = 0, F = 0)$. The fact that we take the presence of oxygen for granted means that the normality ordering makes a world where oxygen is absent quite abnormal. In particular, we require that it satisfy the following properties:

$$(M = 0, O = 1, F = 0) \succeq (M = 1, O = 1, F = 1) \succ (M = 1, O = 0, F = 0).$$

We may or may not consider lit matches to be atypical, although it would be strange to have an ordering where the match being lit is more typical than the match not being lit. After all, a given match can light only once and burns out quickly, so it must be at least fairly typical for it not to be lit. Note that in the partial order, the first world is the witness for $M = 1$ being a cause of $F = 1$, the second is the actual world, and the third is the witness for $O = 1$ being a cause. Thus, we are inclined to judge $M = 1$ a cause and not judge $O = 1$ a cause. More generally, we can understand the setting of an exogenous variable in the HP framework to be one such that a world where it does not have that setting is much more abnormal that a world where it does.

Note that if the fire occurred in a special chamber in a scientific laboratory that is normally voided of oxygen, then we would have a different normality ordering. Now the presence of oxygen is atypical, and the witness for $O = 1$ being a cause is as normal as (or at least not strictly less normal than) the witness for $M = 1$ being a cause. And this corresponds with our intuition that, in such a case, we would be willing to judge the presence of oxygen an actual cause of the fire.

Our treatment of Knobe effects and background conditions is likely to produce a familiar complaint. It is common in discussions of causation to note that while people commonly do make these kinds of discriminations, it is in reality a philosophical mistake to do so. For example, John Stuart Mill writes:

> It is seldom, if ever, between a consequent and a single antecedent, that this invariable sequence subsists. It is usually between a consequent and the sum of several antecedents; the concurrence of all of them being requisite to produce, that is, to be certain of being followed by, the consequent. In such cases it is very common to single out one only of the antecedents under the denomination of Cause, calling the others mere Conditions ... The real cause, is the whole of these antecedents; and we have, philosophically speaking, no right to give the name of cause to one of them, exclusively of the others. [Mill 1856, pp. 360–361].

David Lewis says:

> We sometimes single out one among all the causes of some event and call it “the” cause, as if there were no others. Or we single out a few as the “causes”, calling the rest mere “causal factors” or “causal conditions” . . . I have nothing to say about these principles of invidious discrimination. [Lewis 1973, pp. 558–559]

And Ned Hall adds:

> When delineating the causes of some given event, we typically make what are, from the present perspective, invidious distinctions, ignoring perfectly good causes because they are not sufficiently salient. We say that the lightning bolt caused the forest fire, failing to mention the contribution of the oxygen in the air, or the presence of a sufficient quantity of flammable material. But in the egalitarian sense of “cause,” a complete inventory of the fire’s causes must include the presence of oxygen and of dry wood. [Hall 2004, p. 228]

The concern is that because a cause is not salient, or because it would be inappropriate to assert that it is a cause in some conversational context, we are mistakenly inferring that it is not a cause at all.

The “egalitarian” notion of cause is entirely appropriate at the level of causal structure, as represented by the equations of a causal model. These equations represent objective features of the world, and are not sensitive to factors such as contextual salience. We think that it is a mistake, however, to look for this same objectivity in actual causation. Hitchcock and Knobe (2009) argue that it is in part because of its selectivity that the concept of actual causation earns its keep.

7.5 Causal chains

There has been considerable debate in the philosophical literature over whether causation is transitive, that is, whether whenever A causes B, and B causes C, then A causes C. Lewis [2000], for example, defends the affirmative, while Hitchcock [2001] argues for the negative. But even those philosophers who have denied that causation is transitive in general have not questioned the transitivity of causation in simple causal chains, where the final effect counterfactually depends on the initial cause. By contrast, the law does not assign causal responsibility for sufficiently remote consequences of an action. For example, in Regina v. Faulkner [1877], a well-known Irish case, a lit match aboard a ship caused a cask of rum to ignite, causing the ship to burn, which resulted in a large financial loss by Lloyd’s insurance, leading to the suicide of a financially ruined insurance executive. The executive’s widow sued for compensation, and it was ruled that the negligent lighting of the match was not a cause (in the legally relevant sense) of his death. Moore [2009] uses this type of case to argue that our ordinary notion of actual causation is graded, rather than all-or-nothing, and that it can attenuate over the course of a causal chain.

Our account of actual causation can make sense of this kind of attenuation. We can represent the case of Regina v. Faulkner using a causal model with nine random variables:

- $M = 1$ if the match is lit, 0 if it is not;
- $R = 1$ if there is rum in the vicinity of the match, 0 if not;
- $RI = 1$ if the rum ignites, 0 if it does not;
- $F = 1$ if there is further flammable material near the rum, 0 if not;

- $SD = 1$ if the ship is destroyed, 0 if not;
- $LI = 1$ if the ship is insured by Lloyd's, 0 if not;
- $LL = 1$ if Lloyd's suffers a loss, 0 if not;
- $EU = 1$ if the insurance executive was mentally unstable, 0 if not;
- $ES = 1$ if the executive commits suicide, 0 if not.

There are four structural equations:

$$\begin{aligned}RI &= \min(M, R)\\ SD &= \min(RI, F)\\ LL &= \min(SD, LI)\\ ES &= \min(LL, EU).\end{aligned}$$

This model is shown graphically in Figure 7.5. The exogenous variables are such that M, R, F, LI, and EU are all 1, so in the actual world, all variables take the value 1. Intuitively, the events $M = 1$, $RI = 1$, $SD = 1$, $LL = 1$, and $ES = 1$ form a causal chain. The HP definition rules that the first four events are all actual causes of $ES = 1$.

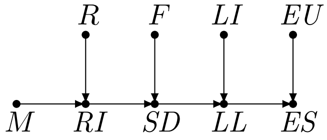

Figure 4: Attenuation in a causal chain.

Let us now assume that, for the variables M, R, F, LI, and EU, 0 is the typical value, and 1 is the atypical value. Thus, our normality ranking assigns a higher rank to worlds where more of these variables take the value 0. For simplicity, consider just the most proximate and the most distal links in the chain: $LL = 1$ and $M = 1$, respectively. The world $(M = 0, R = 1, RI = 0, F = 1, SD = 0, LI = 1, LL = 0, EU = 1, ES = 0)$ is a witness of $M = 1$ being a cause of $ES = 1$. This is quite an abnormal world, although more normal than the actual world, so $M = 1$ does count as an actual cause of $ES = 1$ by our revised definition. Note that if any of the variables R, F, LI, or EU is set to 0, then we no longer have a witness. Intuitively, ES counterfactually depends on M only when all of these other variables take the value 1. Now consider the event $LL = 1$. The world $(M = 0, R = 0, RI = 0, F = 0, SD = 0, LI = 0, LL = 0, EU = 1, ES = 0)$ is a witness for $LL = 1$ being an actual cause of $ES = 1$. This witness is significantly more normal than the best witness for $M = 1$ being a cause. Intuitively, $LL = 1$ needs fewer atypical conditions to be present in order to generate the outcome $ES = 1$. It requires only the instability of the executive, but not the presence of rum, other flammable materials, and so on. Hence, the revised account predicts that we are more strongly inclined to judge that $LL = 1$ is an actual cause of $ES = 1$ than $M = 1$. Nonetheless, the witness for $M = 1$ being a cause is still more normal than the actual world, so we still have some inclination to judge it an actual cause. As Moore [2009] recommends, the revised account yields a graded notion of actual causation.

Note that the extent to which we have attenuation of actual causation over a causal chain is not just a function of spatiotemporal distance or the number of links. It is, rather, a function of how abnormal the circumstances are that must be in place if the causal chain is going to run from start to finish. In the postscript of [Lewis 1986], Lewis uses the phrase "sensitive causation" to describe cases of causation that depend on a complex configuration of background circumstances. For example, he describes a case where he writes a strong letter of recommendation for candidate A, thus earning him a job and displacing second-place candidate B, who then accepts a job at her second choice of institutions, displacing runner-up C, who then accepts a job at another university, where he meets his spouse, and they have a child, who later dies. While Lewis claims that his writing the letter is indeed a cause of the death, it is a highly sensitive cause, requiring an elaborate set of detailed conditions to be present. Woodward [2006] says that such causes are "unstable". Had the circumstances been slightly different, writing the letter would not have produced this effect (either the effect would not have occurred, or it would not have been counterfactually dependent on the letter). Woodward argues that considerations of stability often inform our causal judgments. Our definition allows us to take these considerations into account.

7.6 Legal doctrines of intervening causes

In the law, it is held that one is not causally responsible for some outcome when one's action led to that outcome only via the intervention of a later agent's deliberate action, or some very improbable event. For example, if Anne negligently spills gasoline, and Bob carelessly throws a cigarette in the spill, then Anne's action is a cause of the fire. But if Bob maliciously throws a cigarette in the gas, then Anne is not considered a cause [Hart and Honoré 1985].[13] This reasoning often seems strange to philosophers, but legal theorists find it very natural. As we now show, we can model this judgment in our framework.

In order to fully capture the legal concepts, we need to represent the mental states of the agents. We can do this with the following six variables:

- $AN = 1$ if Anne is negligent, 0 if she isn't;
- $AS = 1$ if Anne spills the gasoline, 0 if she doesn't;
- $BC = 1$ if Bob is careless (i.e. doesn't notice the gasoline), 0 if not;
- $BM = 1$ if Bob is malicious, 0 otherwise;
- $BT = 1$ if Bob throws a cigarette, 0 if he doesn't;
- $F = 1$ if there is a fire, 0 if there isn't.

We have the following equations:

$$\begin{array}{l} F = \min(AS, BT); \\ AS = AN; \\ BT = \max(BC, BM, 1 - AS). \end{array}$$

This model is shown graphically in Figure 7.6. Note that we have made somewhat arbitrary stipulations about what happens in the case where Bob is both malicious and careless, and in the cases where Anne

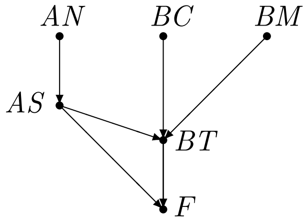

Figure 5: An intervening cause.

does not spill; this is not clear from the usual description of the example. These stipulations do not affect our analysis.

We assume that BM, BC, and AN typically take the value 0. But we can say more. Legally, actions can be graded on a scale of normality. Roughly speaking, in descending order of normality, we have

- prudent and reasonable;
- negligent;
- reckless;
- criminal/intentional.

While our practice so far has been to avoid comparisons of typicality *between* variables, such a comparison clearly seems warranted in this case. Specifically, we have in decreasing order of typicality:

1. $BC = 1$;
2. $AN = 1$;
3. $BM = 1$.

Thus, when we compare the normality of worlds in which two of these variables take the value 0, and one takes the value 1, we would expect the world with $BC = 1$ to be most normal, the world with $AN = 1$ to be next most normal, and the world with $BM = 1$ to be least normal. Consider first the case where Bob is careless. Then in the actual world we have

$$(BM = 0, BC = 1, BT = 1, AN = 1, AS = 1, F = 1).$$

In our structural equations, $F = 1$ depends counterfactually on both $BC = 1$ and $AN = 1$. Thus AC rules that both are actual causes. (We can just take $\vec{W} = \emptyset$ in both cases.) The best witness for $AN = 1$ is

$$(BM = 0, BC = 1, BT = 1, AN = 0, AS = 0, F = 0),$$

while the best witness for $BC = 1$ is

$$(BM = 0, BC = 0, BT = 0, AN = 1, AS = 1, F = 0).$$

[13] This example is based on the facts of Watson v. Kentucky and Indiana Bridge and Railroad [1910].

Both of these worlds are more normal than the actual world. The first is more normal because AN takes the value 0 instead of 1. The second world is more normal than the actual world because BC takes the value 0 instead of 1. Hence, the revised theory judges that both are actual causes. However, the best witness for $AN = 1$ is *more normal* than the best witness for $BC = 1$. The former witness has $BC = 1$ and $AN = 0$, while the latter witness has $BC = 0$ and $AN = 1$. Since $AN = 1$ is more atypical than $BC = 1$, the first witness is more normal. This means that we are more inclined to judge that Anne's negligence is an actual cause of the fire than that Bob's carelessness is.

Now consider the case where Bob is malicious. The actual world is

$$(BM = 1, BC = 0, BT = 1, AN = 1, AS = 1, F = 1).$$

Again, AC straightforwardly rules that both $BM = 1$ and $AN = 1$ are actual causes. The best witness for $AN = 1$ is

$$(BM = 1, BC = 0, BT = 1, AN = 0, AS = 0, F = 0),$$

while the best witness for $BM = 1$ is

$$(BM = 0, BC = 0, BT = 0, AN = 1, AS = 1, F = 0).$$

Again, both of these worlds are more normal than the actual world, so our revised theory judges that both are actual causes. However, now the best witness for $AN = 1$ is *less normal* than the best witness for $BM = 1$, since $BM = 1$ is more atypical than $AN = 1$. So now our theory predicts that we are more inclined to judge that Bob's malice is an actual cause of the fire than that Anne's negligence is.

Recall that in Section 7.3 we saw how judgments of the causal status of the administrative assistant's action changed, depending on the normative status of the professor's action. Something similar is happening here: the causal status of Anne's action changes with the normative status of Bob's action. This example also illustrates how context can play a role in determining what is normal and abnormal. In the legal context, there is a clear ranking of norm violations.

We end this subsection with a technical note. We said that in the case where Bob is inattentive, the best witness for $AN = 1$ being an actual cause of $F = 1$ is $(BM = 0, BC = 1, BT = 1, AN = 0, AS = 0, F = 0)$; whereas in the case where Bob is malicious, the best witness is $(BM = 1, BC = 0, BT = 1, AN = 0, AS = 0, F = 0)$. Since the first of these worlds is more normal than the second, these claims require that the first world *not* be a witness in the case where Bob is malicious ($BM = 1$). This is correct, but assessing it requires paying close attention to the details of clause AC2(b). Consider the case where Bob is malicious; the actual world is $(BM = 1, BC = 0, BT = 1, AN = 1, AS = 1, F = 1)$. Could we, for example, take $\vec{W} = \{BM, BC\}$ and $\vec{w} = \{0, 1\}$, so that our setting is $BM = 0, BC = 1$? This setting satisfies AC2(a). With this setting of BM and BC, if AN were set to 0, F would also be 0. So far, so good. But this setting does not satisfy AC2(b). Take $\vec{W}' = \{BM\} \subseteq \vec{W}$. If we set BM to 0 (the setting it receives in $\vec{w}$), and make no other interventions, then we get $F = 0$. That is, the actual outcome $F = 1$ does not survive when we make the partial setting $\vec{W}' = \{0\}$. AC2(b) then rules that $\vec{w}$ is not a legitimate setting of $\vec{W}'$. As we show in the appendix, however, this result does *not* hold for the earlier version of AC proposed in [Halpern and Pearl 2001].

7.7 Preemption and short circuits

Recall our example of preemption (Example 3.3). in which an assassin poisoned the victim's drink with a backup present. It is convenient to start our exposition with a simplified causal model that omits the variable representing the backup's readiness. (Recall that for purposes of showing the assassin's action to be an actual cause, it did not much matter what we did with that variable.) Thus our variables are:

- $A = 1$ if the assassin poisons the drink, 0 if not;
- $B = 1$ if the backup poisons the drink, 0 if not;
- $D = 1$ if the victim dies, 0 if not.

The equations are

$$\begin{aligned} &B = (1 - A); \\ &D = \max(A, B). \end{aligned}$$

The structure is the same as that depicted in Figure 7.5, with R omitted. In the actual world, $A = 1, B = 0, D = 1$. AC rules that $A = 1$ is an actual cause of $D = 1$. We can let $\vec{W} = \{B\}$, and $\vec{w} = \{0\}$. This satisfies condition AC2. We thus get that the world $(A = 0, B = 0, D = 0)$ is a witness for $A = 1$ being an actual cause of $D = 1$.

On the new account, however, this is not enough. We must also ensure that the witness is sufficiently normal. It seems natural to take $A = 0$ and $B = 0$ to be typical, and $A = 1$ and $B = 1$ to be atypical. It is morally wrong, unlawful, and highly unusual for an assassin to be poisoning one's drink. This gives rise to a normality ranking in which the witness $(A = 0, B = 0, D = 0)$ is more normal than the actual world in which $(A = 1, B = 0, D = 1)$. So the revised account still rules that $A = 1$ is an actual cause of $D = 1$.

Now let us reconsider Example 4.2, in which two bodyguards conspire to make it appear as if one of them has foiled an assassination attempt. Let us model this story using the following variables:

- $A = 1$ if the first bodyguard puts in the antidote, 0 otherwise;
- $P = 1$ if the second bodyguard puts in the poison, 0 otherwise;
- $\mathit{VS} = 1$ if the victim survives, 0 otherwise.

The equations are

$$\begin{aligned} &P = A \\ &\mathit{VS} = \max(A, (1 - P)). \end{aligned}$$

In the actual world, $A = 1, P = 1$, and $\mathit{VS} = 1$. This model is shown graphically in Figure 7.7. A quick inspection reveals that this model is isomorphic to the model for our preemption case, substituting P for $1 - B$ and VS for D. And indeed, AC rules that $A = 1$ is an actual cause of $\mathit{VS} = 1$, with $(A = 0, P = 1, \mathit{VS} = 0)$ as witness. Intuitively, if we hold fixed that the second bodyguard added the poison, the victim would not have survived if the first bodyguard had not added the antidote first. Yet intuitively, many people judge that the first bodyguard's adding the antidote is not a cause of survival (or a preventer of death), since the only threat to victim's life was itself caused by that very same action.

In order to capture the judgment that $A = 1$ is not an actual cause of $\mathit{VS} = 1$, we must appeal to the normality ranking. Here we have to be careful. Suppose that we decide that the typical value of both A and P is 0, and the atypical value 1. This would seem *prima facie* reasonable. This would give us a normality ordering in which the witness $(A = 0, P = 1, \mathit{VS} = 0)$ is more normal than the actual world $(A = 1, P = 1, \mathit{VS} = 1)$. Intuitively, the value of A is more typical in the first world, and the value of

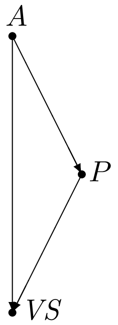

Figure 6: The poisoning example reconsidered.

P is the same in both words, so the first world is more normal overall than the second. If we reason this way, the modified theory still makes $A = 1$ an actual cause of $VS = 1$.

There is, however, a subtle mistake in this way of thinking about the normality ordering. The variable P is not fixed by exogenous variables that are outside of the model; it is also determined by A, according to the equation $P = A$. This means that when we think about what is typical for P, we should rank not just the typicality of particular values of P, but the typicality of different ways for P to depend on A. And the most natural ranking is the following, in decreasing order of typicality:

1. $P = 0$, regardless of A;

2. $P = A$;

3. $P = 1$ regardless of A.

That is, the most typical situation is one where there is no possibility of poison going into the drink. The next most typical is the one that actually occurred. Least typical is one where the second bodyguard puts in poison even when no antidote is added. Put another way, the most typical situation is one where the second bodyguard has no bad intentions, next most typical is where he has deceitful intentions, and least typical is where he has homicidal intentions. This is similar to the ranking we saw in the previous example. Therefore, in the witness world, where $A = 0$ and $P = 1$, P depends on A in a less typical way than in the actual world, in which $A = 1$ and $P = 1$. On this interpretation, the witness world $(A = 0, P = 1, VS = 0)$ is no longer more normal than the actual world $(A = 1, P = 1, VS = 0)$. In fact, using the guidelines we have been following, these two worlds are incomparable.

Thus, while the pattern of counterfactual dependence is isomorphic in the two cases, the normality ranking on worlds is not. The result is that in the case of simple preemption, the witness is more normal than the actual world, while in the second case it is not. This explains the difference in our judgments of actual causation in the two examples. [14]

This example raises an interesting technical point. Our treatment of this example involves an exception to the default assumption that it is typical for endogenous variables to depend on one another in accordance with the structural equations. (Observant readers will note that our treatment of the preemption example earlier in this section did as well.) We can avoid this by adding an additional variable to

[14] Hall [2007, Sections 3.3 and 3.4] presents several examples of what he calls "short circuits". These can be handled in essentially the same way.

the model, akin to the variable representing the backup's readiness in our discussion of the preemption example (Example 3.3). Let I represent the intentions of the second bodyguard, as follows:

- $I = 0$ if he has benign intentions;
- $I = 1$ if he has deceitful intentions;
- $I = 2$ if he has murderous intentions.

Now we can rewrite the equations as follows:

$$\begin{array}{l} P = f(A, I); \\ VS = \max(A, (1 - P)), \end{array}$$

where $f(A, I)$ is the function of A and I that takes the value 0, A, or 1 when $I = 0$, 1, or 2, respectively. This model is shown graphically in Figure 7.7. In this model, we can retain our guideline that endogenous variables typically obey the structural equations. That is, the typicality or atypicality of the way in which P depends on A can be coded by the typicality or atypicality of the values of I. Assume that atypicality increases with increasing values of I. Now the best witnesses for $A = 1$ being an actual cause of $VS = 1$ are $(A = 0, I = 0, P = 1, VS = 0)$ and $(A = 0, I = 2, P = 1, VS = 0)$. The first of these involves P behaving less typically than in the actual world (since it now violates the equation $P = f(A, I)$). The second witness involves I taking a less typical value (2 instead of 1). In the second witness, P is behaving typically: it satisfies the equation $P = f(A, I)$. The abnormality of the situation is now attributed to I. Neither of these witnesses would be as normal as the actual world, so our revised theory would still rule that $A = 1$ is not an actual cause of $VS = 1$.

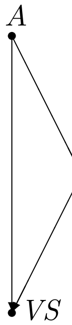

Figure 7: The poisoning example, with intentions.

This raises an interesting technical question. Is it always possible to perform this kind of maneuver? That is, if we have a theory of typicality in which one variable typically violates one of the structural equations, is it always possible, through the addition of suitable variables, to reproduce the normality ordering on worlds using a theory of typicality in which the structural equations are typically obeyed? We conjecture that it is always possible to formally construct additional variables to do this; but there is no guarantee that such variables will correspond to any causal variables that are actually present in a particular situation.

8 Conclusion

According to a naïve counterfactual theory of causation, effects are counterfactually dependent on their causes. This theory faces a number of familiar difficulties. The HP definition of actual causation deals with these by allowing effects to depend on their causes under various contingencies. This definition is too permissive. We have attempted to improve the HP definition by bringing to bear considerations of the normality of contingencies. The more normal the contingency in which the cause affects the outcome, the stronger our inclination to judge that something is a cause. As we have shown, this approach allows us to account for a wide range of judgments of actual causation.

A Appendix

A.1 The original HP definition

Halpern and Pearl [2001] offered an earlier version of Definition AC, which incorporated a more lenient version of clause AC2(b). We reproduce here the full definition from this earlier paper:

Definition A.1: (AC$'$) $\vec{X} = \vec{x}$ 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 $\vec{W}$ 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 $Z \in \vec{Z}$, then both of the following conditions hold:

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

(b$'$) $(M, \vec{u}) \models [\vec{X} = \vec{x}, \vec{W} = \vec{w}, \vec{Z}' = \vec{z}^*]\varphi$ for all subsets $\vec{Z}'$ of $\vec{Z}$.

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

This is the same as Definition 3.1 above, except for clause (b) of AC2, which we have here marked with a prime. AC2(b$'$) requires that φ continue to hold when we set $\vec{W} = \vec{w}$ (and set other variables as well), but it does not require that φ continue to hold when we set arbitrary subsets of $\vec{W}$ to the values assigned in $\vec{w}$. Thus AC$'$ clearly allows more settings $\vec{W} = \vec{w}$ to be permitted.

Halpern and Pearl [2005] rejected AC$'$ in favor of AC, because of a counterexample due to Hopkins and Pearl [2003].

Example A.2: A prisoner dies ($D = 1$) if A loads B's gun ($A = 1$) and B shoots it ($B = 1$), or if C loads her gun and shoots it ($C = 1$). The equation for D is

$$D = \max(\min(A, B), C).$$

In fact, A loads B's gun, B does not shoot, but C does shoot and the prisoner dies ($A = 1, B = 0, C = 1, D = 1$).

AC$'$ rules that $A = 1$ is an actual cause of $D = 1$: even though B did not shoot, A's loading the gun is an actual cause of the prisoner's death. To see why, let $\vec{W} = \{B, C\}$ and $\vec{w} = \{1, 0\}$. In the contingency where B shoots and C doesn't, the prisoner's death is counterfactually dependent on S's action. AC2(b$'$) permits this setting, since setting B to 1 and C to 0 doesn't change the value of D. AC2(b) does not permit this setting, for it would force us to also consider the case where we just set C to 0, in which case $D = 0$. The verdict of AC$'$ clearly seems wrong. ∎

However, we could retain AC$'$, and deal with this case in the same way we dealt with bogus prevention in Section 7.1. The witness for $A = 1$ being an actual cause of $D = 1$ is $(A = 0, B = 1, C = 0, D = 0)$. This is not more normal than the actual world if we make the assumption that B typically takes the value 0. Alternately, we could add a variable for B's bullet in flight, and the structure is similar to a case of preemption. Hence, our revised approach gives us the option of retaining AC$'$ instead of AC.

AC$'$ has some advantages over AC, and some disadvantages. One disadvantage is that the analysis of intervening causes in Section 7.6 does not go through on AC$'$. Consider the case where Bob maliciously throws his cigarette into the gasoline: $(BM = 1, BC = 0, BT = 1, AN = 1, AS = 1, F = 1)$. Under AC, the best witness for $AN = 1$ to be a cause of $F = 1$ was $(BM = 1, BC = 0, BT = 1, AN = 0, AS = 0, F = 0)$. However, under AC$'$, $(BM = 0, BC = 1, BT = 1, AN = 0, AS = 0, F = 0)$ is also a witness. We can take $\vec{W} = \{BM, BC\}$ and $\vec{w} = \{0, 1\}$. Setting BM to 0 and BC to 1 does not change the value of F. AC$'$ does not force us to consider the partial setting $BM = 0$, which would change the value of F. Since this new witness is more normal than the previous witness, it is now the best witness. Since this is the same as the best witness for the case where Bob carelessly throws the cigarette, our account cannot treat the two cases differently.

On the other hand, AC$'$ allows us to provide an elegant account of *double prevention*, as we now show.

A.2 Double prevention

"Double prevention" is a phrase coined by Hall [2004] to describe a case in which a potential preventer of some outcome is itself prevented. Schaffer [2000, 2004] describes the same sort of case, using the term "disconnection". Otte [1986] provides an earlier example. Suppose that a vacationer owns a summer house in the mountains. During the summer, she builds a wall to protect the house from avalanches. While she is away, a vandal destroys the wall. That winter, there is heavy snow, and an avalanche damages her house. The question is whether the vandal's action is an actual cause of the house being damaged. On the one hand, the house would not have been damaged if it weren't for the vandal's action. On the other hand, the vandal did not actually do anything to the house, or initiate a process (such as lighting a fire in the yard) that would damage the house.

Philosophers have been split on this issue, staking out positions similar to those on the issue of causation by omission. We do not review all of these alternatives in detail. Walsh and Sloman [2005] compared the responses of subjects in cases of double prevention with their responses in cases of direct causation, and also in cases of causal irrelevance, and found that subjects gave intermediate ratings in the cases of double prevention. We now show how our account can capture this kind of intermediate judgment.

We use a causal model, with five variables:

- $WB = 1$ if the wall is built, 0 if not;

- $VD = 1$ if the vandal destroys the wall, 0 if not;
- $WW = 1$ if the wall is intact during the winter, 0 if not;
- $A = 1$ if there is an avalanche, 0 if not;
- $HD = 1$ if the house is damaged, 0 if not.

There are two structural equations:

$$\begin{array}{c} WW = \min(WB, 1 - VD) \\ HD = \min(A, 1 - WW). \end{array}$$

This model is represented graphically in Figure A.2. The exogenous variables are such that, in the actual world, $WB = 1$, $VD = 1$, $WW = 0$, $A = 1$, $HD = 1$. This is a case of double prevention, since the wall would have prevented the avalanche from harming the house if it had been intact, but it was prevented from doing so by the vandal. There is a causal chain from VD to WW to HD, but unlike the causal chains discussed in Section 7.5, one link in this chain involves an absence: the absence of the protective wall during the winter. It is for this reason that some writers have classified double prevention together with cases of causation by omission. Since $HD = 1$ depends counterfactually on both $A = 1$ and $VD = 1$, the HP definition rules that both the avalanche and the vandal are actual causes of the damage to the house.

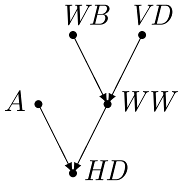

Figure 8: Double prevention.

Now suppose that we add normality to the picture. We assume that the typical values for WB, VD, and A are all 0. This gives rise to a normality ranking in which

$$\begin{array}{ll} & (WB = 0, VD = 0, WW = 0, A = 0, HD = 0) \\ \succ & (WB = 1, VD = 1, WW = 0, A = 0, HD = 0), (WB = 1, VD = 0, WW = 1, A = 1, HD = 0) \\ \succ & (WB = 1, VD = 1, WW = 0, A = 1, HD = 1). \end{array}$$

We can leave open the relative normality of the worlds $(WB = 1, VD = 1, WW = 0, A = 0, HD = 0)$ and $(WB = 1, VD = 0, WW = 1, A = 1, HD = 0)$. The latter is a witness for $VD = 1$ being an actual cause of $HD = 1$; in fact, it is the only witness. The former world is a witness for $A = 1$ being an actual cause. But under AC$'$, $A = 1$ has another, better (more normal) witness, namely $(WB = 0, VD = 0, WW = 0, A = 0, HD = 0)$. The world $(WB = 0, VD = 0, WW = 0, A = 0, HD = 0)$ is a witness for $A = 1$ being a cause of $HD = 1$ because we can take $\vec{W} = \{WB, VD\}$ and $\vec{w} = \{0, 0\}$. Setting $WB = 0$ and $VD = 0$ doesn't change the outcome ($HD = 1$), even if other variables are set at their actual values. AC2(b$'$) doesn't force us to consider any other settings of $\vec{W} = \{WB, VD\}$. Thus, according to the extended version of AC$'$, we should be more inclined to judge that the avalanche is an actual cause of the damage than that the vandal's action caused the damage. The idea here is similar to that discussed in Section 7.5. The vandal's action only makes a difference to the house in the special case where a wall is built, and there is an avalanche. By contrast, the avalanche would make a difference even in the more normal case where both wall and vandal are absent.

AC2(b), however, requires us to also consider interventions that set only some of the variables in $\vec{W}$. When we set $VD = 0$ (and intervene on no other variables), the value of HD does change from 1 to 0. That is, if we intervene only to stop the vandal, the wall gets built, remains intact, and protects the house from the avalanche. Thus AC2(b) rules that that $\{0, 0\}$ is not a legitimate setting for $\vec{W} = \{WB, VD\}$. So under AC, the best witness for $A = 1$ being an actual cause is $(WB = 1, VD = 1, WW = 0, A = 0, HD = 0)$. Under some circumstances, we may be willing to judge that this world is more normal than $(WB = 1, VD = 0, WW = 1, A = 1, HD = 0)$, the best witness for $VD = 1$. For example, if we judge that avalanches are much more atypical than acts of vandalism, we might order the worlds in this way. But the case no longer seems clear cut.

We leave it an open question whether AC or AC$'$ is superior overall.

References

Adams, E. (1975). *The Logic of Conditionals*. Dordrecht, Netherlands: Reidel.

Alicke, M. (1992). Culpable causation. *Journal of Personality and Social Psychology 63*, 368–378.

Beebee, N. (2004). Causing and nothingness. In J. Collins, N. Hall, and L. A. Paul (Eds.), *Causation and Counterfactuals*, pp. 291–308. Cambridge, Mass.: MIT Press.

Cushman, F., J. Knobe, and W. Sinnott-Armstrong (2008). Moral appraisals affect doing/allowing judgments. *Cognition 108*(1), 281–289.

Dowe, P. (2000). *Physical Causation*. Cambridge, U.K.: Cambridge University Press.

Dubois, D. and H. Prade (1991). Possibilistic logic, preferential models, non-monotonicity and related issues. In *Proc. Twelfth International Joint Conference on Artificial Intelligence (IJCAI '91)*, pp. 419–424.

Geffner, H. (1992). High probabilities, model preference and default arguments. *Mind and Machines 2*, 51–70.

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

Goldszmidt, M. and J. Pearl (1992). Rank-based systems: A simple approach to belief revision, belief update and reasoning about evidence and actions. In *Principles of Knowledge Representation and Reasoning: Proc. Third International Conference (KR '92)*, pp. 661–672.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, and L. A. Paul (Eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT Press.

Hall, N. (2007). Structural equations and causation. *Philosophical Studies 132*, 109–136.

Halpern, J. Y. (2008). Defaults and normality in causal structures. In *Principles of Knowledge Representation and Reasoning: Proc. Eleventh International Conference (KR '08)*, pp. 198–208.

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

Halpern, J. Y. and C. Hitchcock (2012). Compact representations of causal models. Unpublished manuscript.

Halpern, J. Y. and J. Pearl (2001). Causes and explanations: A structural-model approach. Part I: Causes. In *Proc. Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI 2001)*, pp. 194–202.

Halpern, J. Y. and J. Pearl (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for Philosophy of Science 56*(4), 843–887.

Hart, H. L. A. and T. Honoré (1985). *Causation in the Law* (second ed.). Oxford, U.K.: Oxford University Press.

Hiddleston, E. (2005). Causal powers. *British Journal for Philosophy of Science 56*, 27–59.

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.

Hitchcock, C. and J. Knobe (2009). Cause and norm. *Journal of Philosophy 106*, 587–612.

Hopkins, M. and J. Pearl (2003). Clarifying the usage of structural models for commonsense causal reasoning. In *Proc. AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Reprinted by Open Court Press, LaSalle, IL, 1958.

Kahneman, D. and D. T. Miller (1986). Norm theory: comparing reality to its alternatives. *Psychological Review 94*(2), 136–153.

Kahneman, D. and A. Tversky (1982). The simulation heuristic. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment Under Incertainty: Heuristics and Biases*, pp. 201–210. Cambridge/New York: Cambridge University Press.

Knobe, J. and B. Fraser (2008). Causal judgment and moral judgment: two experiments. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality*, pp. 441–447. Cambridge, MA: MIT Press.

Kraus, S., D. Lehmann, and M. Magidor (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence 44*, 167–207.

Lewis, D. (1973). Causation. *Journal of Philosophy 70*, 556–567. Reprinted with added "Postscripts" in D. Lewis, *Philosophical Papers*, Volume II, Oxford University Press, 1986, pp. 159–213.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs 13*, 455–476. Reprinted in D. Lewis, *Philosophical Papers*, Volume II, Oxford University Press, 1986, pp. 32–52.

Lewis, D. (1986). Causation. In *Philosophical Papers*, Volume II, pp. 159–213. New York: Oxford University Press. The original version of this paper, without numerous postscripts, appeared in the *Journal of Philosophy* **70**, 1973, pp. 113–126.

Lewis, D. (2000). Causation as influence. *Journal of Philosophy XCVII*(4), 182–197.

Lewis, D. (2004). Void and object. In J. Collins, N. Hall, and L. A. Paul (Eds.), *Causation and Counterfactuals*, pp. 277–290. Cambridge, Mass.: MIT Press.

Marek, W. and M. Truszczyński (1993). *Nonmonotonic Logic*. Berlin/New York: Springer-Verlag.

McGrath, S. (2005). Causation by omission. *Philosophical Studies 123*, 125–148.

Mill, J. S. (1856). *A System of Logic, Ratiocinative and Inductive* (Fourth ed.). London: John W. Parker and Son.

Moore, M. S. (2009). *Causation and Responsibility*. Oxford, UK: Oxford University Press.

Otte, R. (1986). Reichenbach, causation, and explanation. In A. I. Fine and P. K. Machamer (Eds.), *PSA 1986 Volume I*, pp. 59–65. East Lansing, Michigan: Philosophy of Science Association.

Pearl, J. (1989). Probabilistic semantics for nonmonotonic reasoning: a survey. In *Proc. First International Conference on Principles of Knowledge Representation and Reasoning (KR '89)*, pp. 505–516. Reprinted in G. Shafer and J. Pearl (Eds.), *Readings in Uncertain Reasoning*, pp. 699–710. San Francisco: Morgan Kaufmann, 1990.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Regina v. Faulkner (1877). Court of Crown Cases Reserved, Ireland). In *13 Cox Criminal Cases 550*.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence 13*, 81–132.

Reiter, R. (1987). Nonmonotonic reasoning. In J. F. Traub, B. J. Grosz, B. W. Lampson, and N. J. Nilsson (Eds.), *Annual Review of Computer Science, Volume 2*, pp. 147–186. Palo Alto, Calif.: Annual Reviews Inc.

Schaffer, J. (2000). Causation by disconnection. *Philosophy of Science 67*, 285–300.

Schaffer, J. (2004). Causes need not be physically connected to their effects. In C. Hitchcock (Ed.), *Contemporary Debates in Philosophy of Science*, pp. 197–216. Oxford, U.K.: Basil Blackwell.

Shoham, Y. (1987). A semantical approach to nonmonotonic logics. In *Proc. 2nd IEEE Symposium on Logic in Computer Science*, pp. 275–279. Reprinted in M. L. Ginsberg (Ed.), *Readings in Nonmonotonic Reasoning*, pp. 227–250. San Francisco: Morgan Kaufman, 1987.

Spohn, W. (1988). Ordinal conditional functions: a dynamic theory of epistemic states. In W. Harper and B. Skyrms (Eds.), *Causation in Decision, Belief Change, and Statistics*, Volume 2, pp. 105–134. Dordrecht, Netherlands: Reidel.

Sytsma, J., J. Livengood, and D. Rose (2010). Two types of typicality: Rethinking the role of statistical typicality in folk causal attributions. Unpublished manuscript.

Tversky, A. and D. Kahneman (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology 5*, 207–232.

Walsh, C. R. and S. A. Sloman (2005). The meaning of cause and prevent: The role of causal mechanism. In *Proc. 27th Annual Conference of the Cognitive Science Society*, pp. 2311–2316.

Watson v. Kentucky and Indiana Bridge and Railroad (1910). 137 Kentucky 619. In *126 SW 146*.

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

Woodward, J. (2006). Sensitive and insensitive causation. *Philosophical Review 115*, 1–50.

Wright, R. W. (1985). Causation in tort law. *California Law Review 73*, 1735–1828.