23

Actual Causation and the Art of Modeling

Joseph Y. Halpern and Christopher Hitchcock

1 Introduction

In *The Graduate*, Benjamin Braddock (Dustin Hoffman) is told that the future can be summed up in one word: "Plastics". One of us (Halpern) recalls that in roughly 1990, Judea Pearl told him that the future was in causality. Pearl's own research was largely focused on causality in the years after that; his seminal contributions are widely known. We were among the many influenced by his work. We discuss one aspect of it, *actual causation*, in this article, although a number of our comments apply to causal modeling more generally.

Pearl introduced a novel account of actual causation in Chapter 10 of *Causality*, which was later revised in collaboration with one of us [Halpern and Pearl 2005]. In some ways, Pearl's approach to actual causation can be seen as a contribution to the philosophical project of trying to analyze actual causation in terms of counterfactuals, a project associated most strongly with David Lewis [1973a]. But Pearl's account was novel in at least two important ways. The first was his use of *structural equations* as a tool for modeling causality. In the philosophical literature, causal structures were often represented using so-called *neuron diagrams*, but these are not (and were never intended to be) all-purpose representational tools. (See [Hitchcock 2007b] for a detailed discussion of the limitations of neuron diagrams.) We believe that the lack of a more adequate representational tool had been a serious obstacle to progress. Second, while the philosophical literature on causality has focused almost exclusively on actual causality, for Pearl, actual causation was a rather specialized topic within the study of causation, peripheral to many issues involving causal reasoning and inference. Thus, Pearl's work placed the study of actual causation within a much broader context.

The use of structural equations as a model for causal relationships was well known long before Pearl came on the scene; it seems to go back to the work of Sewall Wright in the 1920s (see [Goldberger 1972] for a discussion). However, the details of the framework that have proved so influential are due to Pearl. Besides the Halpern-Pearl approach mentioned above, there have been a number of other closely-related approaches for using structural equations to model actual causation; see, for example, [Glymour and Wimberly 2007; Hall 2007; Hitchcock 2001; Hitchcock 2007a; Woodward 2003]. The goal of this paper is to look more carefully at the modeling of causality using structural equations. For definiteness, we use the

Halpern-Pearl (HP) version [Halpern and Pearl 2005] here, but our comments apply equally well to the other variants.

It is clear that the structural equations can have a major impact on the conclusions we draw about causality—it is the equations that allow us to conclude that lower air pressure is the cause of the lower barometer reading, and not the other way around; increasing the barometer reading will not result in higher air pressure. The structural equations express the effects of *interventions*: what happens to the bottle if it is hit with a hammer; what happens to a patient if she is treated with a high dose of the drug, and so on. These effects are, in principle, objective; the structural equations can be viewed as describing objective features of the world. However, as pointed out by Halpern and Pearl [2005] and reiterated by others [Hall 2007; Hitchcock 2001; Hitchcock 2007a], the choice of variables and their values can also have a significant impact on causality. Moreover, these choices are, to some extent, subjective. This, in turn, means that judgments of actual causation are subjective.

Our view of actual causation being at least partly subjective stands in contrast to the prevailing view in the philosophy literature, where the assumption is that the job of the philosopher is to analyze the (objective) notion of causation, rather like that of a chemist analyzing the structure of a molecule. This may stem, at least in part, from failing to appreciate one of Pearl's lessons: actual causality is only part of the bigger picture of causality. There can be an element of subjectivity in ascriptions of actual causality without causation itself being completely subjective. In any case, the experimental evidence certainly suggests that people's views of causality are subjective, even when there is no disagreement about the relevant structural equations. For example, a number of experiments show that broadly normative considerations, including the subject's own moral beliefs, affect causal judgment. (See, for example, [Alicke 1992; Cushman 2009; Cushman, Knobe, and Sinnott-Armstrong 2008; Hitchcock and Knobe 2009; Knobe and Fraser 2008].) Even in relatively non-controversial cases, people may want to focus on different aspects of a problem, and thus give different answers to questions about causality. For example, suppose that we ask for the cause of a serious traffic accident. A traffic engineer might say that the bad road design was the cause; an educator might focus on poor driver's education; a sociologist might point to the pub near the highway where the driver got drunk; a psychologist might say that the cause is the driver's recent breakup with his girlfriend.[1] Each of these answers is reasonable. By appropriately choosing the variables, the structural equations framework can accommodate them all.

Note that we said above "by appropriately choosing the variables". An obvious question is "What counts as an appropriate choice?". More generally, what makes a model an appropriate model? While we do want to allow for subjectivity, we need

[1] This is a variant of an example originally due to Hanson [1958].

to be able to justify the modeling choices made. A lawyer in court trying to argue that faulty brakes were the cause of the accident needs to be able to justify his model; similarly, his opponent will need to understand what counts as a legitimate attack on the model. In this paper we discuss what we believe are reasonable bases for such justifications. Issues such as model stability and interactions between the events corresponding to variables turn out to be important.

Another focus of the paper is the use of defaults in causal reasoning. As we hinted above, the basic structural equations model does not seem to suffice to completely capture all aspects of causal reasoning. To explain why, we need to briefly outline how actual causality is defined in the structural equations framework. Like many other definitions of causality (see, for example, [Hume 1739; Lewis 1973b]), the HP definition is based on counterfactual dependence. Roughly speaking, A is a cause of B if, had A not happened (this is the counterfactual condition, since A did in fact happen) then B would not have happened. As is well known, this naive definition does not capture all the subtleties involved with causality. Consider the following example (due to Hall [2004]): Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle had Suzy not thrown. Thus, according to the naive counterfactual definition, Suzy's throw is not a cause of the bottle shattering. This certainly seems counterintuitive.

The HP definition deals with this problem by taking A to be a cause of B if B counterfactually depends on A *under some contingency*. For example, Suzy's throw is the cause of the bottle shattering because the bottle shattering counterfactually depends on Suzy's throw, under the contingency that Billy doesn't throw. (As we will see below, there are further subtleties in the definition that guarantee that, if things are modeled appropriately, Billy's throw is not also a cause.)

While the definition of actual causation in terms of structural equations has been successful at dealing with many of the problems of causality, examples of Hall [2007], Hiddleston [2005], and Hitchcock [2007a] show that it gives inappropriate answers in cases that have structural equations isomorphic to ones where it arguably gives the appropriate answer. This means that, no matter how we define actual causality in the structural-equations framework, the definition must involve more than just the structural equations. Recently, Hall [2007], Halpern [2008], and Hitchcock [2007a] have suggested that using defaults might be a way of dealing with the problem. As the psychologists Kahneman and Miller [1986, p. 143] observe, "an event is more likely to be undone by altering exceptional than routine aspects of the causal chain that led to it". This intuition is also present in the legal literature. Hart and Honoré [1985] observe that the statement "It was the presence of oxygen that caused the fire" makes sense only if there were reasons to view the presence of oxygen as abnormal.

As shown by Halpern [2008], we can model this intuition formally by combining a well-known approach to modeling defaults and normality, due to Kraus, Lehmann,

and Magidor [1990] with the structural-equation model. Moreover, doing this leads to a straightforward solution to the problem above. The idea is that, when showing that if A hadn't happened then B would not have happened, we consider only contingencies that are at least as normal as the actual world. For example, if someone typically leaves work at 5:30 PM and arrives home at 6, but, due to unusually bad traffic, arrives home at 6:10, the bad traffic is typically viewed as the cause of his being late, not the fact that he left at 5:30 (rather than 5:20).

But once we add defaults to the model, the problem of justifying the model becomes even more acute. We not only have to justify the structural equations and the choice of variables, but also the default theory. The problem is exacerbated by the fact that default and "normality" have a number of interpretations. Among other things, they can represent moral obligations, societal conventions, prototypicality information, and statistical information. All of these interpretations are relevant to understanding causality; this makes justifying default choices somewhat subtle.

The rest of this paper is organized as follows. In Sections 2 and 3, we review the notion of causal model and the HP definition of actual cause; most of this material is taken from [Halpern and Pearl 2005]. In Section 4, we discuss some issues involved in the choice of variables in a model. In Section 5, we review the approach of [Halpern 2008] for adding considerations of normality to the HP framework, and discuss some modeling issues that arise when we do so. We conclude in Section 6.

2 Causal Models

In this section, we briefly review the HP definition of causality. The description of causal models given here is taken from [Halpern 2008], which in turn is based on that of [Halpern and Pearl 2005].

The HP approach assumes that the world is described in terms of 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 take the world to be described by three random variables:

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

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 either a match is lit or lightning strikes, then a fire starts, we could use the random variables ML, F, and L as above, with the equation $F = \max(L, ML)$. (Alternately, if a fire requires both causes to be present, the equation for F becomes $F = \min(L, ML)$.) The equality sign in this equation should be thought of more like an assignment statement in programming languages; once we set the values of F

and L, then the value of F is set to their maximum. However, despite the equality, if a forest fire starts some other way, that does not force the value of either ML or L to be 1.

It is conceptually useful to split the random variables into two sets: the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. For example, in the forest-fire example, the variables ML, L, and F 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. These factors are all determined by the exogenous variables.

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*). $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$. This mathematical notation just makes precise the fact that F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. If there is one exogenous variable U and three endogenous variables, X, Y, and Z, then F_X defines the values of X in terms of the values of Y, Z, and U. For example, we might have $F_X(u, y, z) = u + y$, which is usually written as $X \leftarrow U + Y$.[2] Thus, if $Y = 3$ and $U = 2$, then $X = 5$, regardless of how Z is set.

In the running forest-fire example, suppose that we have an exogenous random variable U that determines the values of L and ML. Thus, U has four possible values of the form (i, j), where both of i and j are either 0 or 1. The i value determines the value of L and the j value determines the value of ML. Although F_L gets as arguments the vale of U, ML, and F, in fact, it depends only on the (first component of) the value of U; that is, $F_L((i, j), m, f) = i$. Similarly, $F_{ML}((i, j), l, f) = j$. The value of F depends only on the value of L and ML. *How* it depends on them depends on whether either cause by itself is sufficient for the forest fire or whether both are necessary. If either one suffices, then $F_F((i, j), l, m) = \max(l, m)$, or, perhaps more comprehensibly, $F = \max(L, ML)$; if both are needed, then $F = \min(L, ML)$. For future reference, call the former model the *disjunctive* model, and the latter the *conjunctive* model.

The key role of the structural equations is to define what happens in the presence of external interventions. For example, we can explain what happens if the arsonist

[2]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.

does *not* drop the match. In the disjunctive model, there is a forest fire 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 $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model denoted $M_{X \leftarrow x}$. In the new causal model, the equation for X is very simple: X is just set to x; the remaining equations are unchanged. More formally, $M_{X \leftarrow x} = (\mathcal{S}, \mathcal{F}^{X \leftarrow x})$, where $\mathcal{F}^{X \leftarrow x}$ is the result of replacing the equation for X in $\mathcal{F}$ by $X = x$.

The structural equations describe *objective* information about the results of interventions, that can, in principle, be checked. Once the modeler has selected a set of variables to include in the model, *the world* determines which equations among those variables correctly represent the effects of interventions.[3] By contrast, the *choice* of variables is subjective; in general, there need be no objectively "right" set of exogenous and endogenous variables to use in modeling a problem. We return to this issue in Section 4.

It may seem somewhat circular to use causal models, which clearly already encode causal information, to define actual causation. Nevertheless, as we shall see, there is no circularity. The equations of a causal model do not represent relations of *actual causation*, the very concept that we are using them to define. Rather, the equations characterize the results of *all possible* interventions (or at any rate, all of the interventions that can be represented in the model) without regard to what actually happened. Specifically, the equations do not depend upon the actual values realized by the variables. For example, the equation $F = \max(L, ML)$, by itself, does not say anything about whether the forest fire was actually caused by lightning or by an arsonist, or, for that matter, whether a fire even occurred. By contrast, relations of actual causation depend crucially on how things actually play out.

A sequence of endogenous $X_1, \ldots, X_n$ of is a *directed path* from X_1 to X_n if the value of X_{i+1} (as given by $F_{X_{i+1}}$) depends on the value of X_i, for $1 = 1, \ldots, n-1$. In this paper, following HP, we restrict our discussion to *acyclic* causal models, where causal influence can be represented by an acyclic Bayesian network. That is, there is no cycle $X_1, \ldots, X_n, X_1$ of endogenous variables that forms a directed path from X_1 to itself. 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.

[3] In general, there may be uncertainty about the causal model, as well as about the true setting of the exogenous variables in a causal model. Thus, we may be uncertain about whether smoking causes cancer (this represents uncertainty about the causal model) and uncertain about whether a particular patient actually smoked (this is uncertainty about the value of the exogenous variable that determines whether the patient smokes). This uncertainty can be described by putting a probability on causal models and on the values of the exogenous variables. We can then talk about the probability that A is a cause of B.

3 The HP Definition of Actual Cause

3.1 A language for describing causes

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

A causal formula ψ is true or false in a causal model, given a context. As usual, we write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in causal model M given context $\vec{u}$. The $\models$ relation is defined inductively. $(M, \vec{u}) \models X = x$ if the variable X has value x in the unique (since we are dealing with acyclic models) solution to the equations in M in context $\vec{u}$ (that is, the unique vector of values for the 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, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\phi$ if $(M_{\vec{Y} \leftarrow \vec{y}}, \vec{u}) \models \phi$. We write $M \models \phi$ if $(M, \vec{u}) \models \phi$ for all contexts $\vec{u}$.

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

3.2 A preliminary definition of causality

The HP definition of causality, like many others, is based on counterfactuals. The idea is that if A and B both occur, then A is a cause of B if, if A hadn't occurred, 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, ..., where, 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.

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 causes of the fire. (In general, there may be more than one actual cause of an outcome.)

A more subtle problem is what philosophers have called *preemption*, which is illustrated by the rock-throwing example from the introduction. As we observed, according to a naive counterfactual definition of causality, Suzy's throw would not be a cause.

The HP definition deals with the first problem by defining causality as counterfactual dependency *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 dropping of the match under the contingency that the lightning does not strike.

Unfortunately, we cannot use this simple solution to treat the case of preemption. We do not want to make Billy's throw the cause of the bottle shattering by considering the contingency that Suzy does not throw. So if our account is to yield the correct verdict in this case, it will be necessary to limit the contingencies that can be considered. The reason that we consider Suzy's throw to be the cause and Billy's throw not to be the cause is that Suzy's rock hit the bottle, while Billy's did not. Somehow the definition of actual cause must capture this obvious intuition.

With this background, we now give the preliminary version of the HP definition of causality. Although the definition is labeled "preliminary", it is quite close to the final definition, which is given in Section 5. 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 $X_1 = x_1 \wedge \ldots \wedge X_k = x_k$—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; this is not equivalent to "either A is a cause of B or A is a cause of B'".

DEFINITION 1. (Actual cause; preliminary version) [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 $(M, \vec{u}) \models \phi$.

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

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

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, consisting of one or more directed paths from variables in $\vec{X}$ to variables in ϕ. Intuitively, changing the value(s) of some variable(s) in $\vec{X}$ results in changing the value(s) of some variable(s) in $\vec{Z}$, which results in the value(s) 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 ϕ would have been false. However, this effect of the value of $\vec{X}$ on the truth value of ϕ may not hold in the actual context; the value of $\vec{W}$ may have to be different to allow this effect to manifest itself. For example, consider the context 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. Essentially, it ensures that $\vec{X}$ alone suffices to bring about the change from ϕ to $\neg\phi$; 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, when $\vec{X} = \vec{x}$, although the values of variables on the causal path (i.e., the variables $\vec{Z}$) may be perturbed by the change to $\vec{W}$, this perturbation has no impact on the value of ϕ. If $(M, \vec{u}) \models \vec{Z} = \vec{z}^*$, then z^* is the value of the variable Z in the context $\vec{u}$. 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 $\vec{X} = \vec{x}$.

EXAMPLE 2. For the forest-fire example, let M be the disjunctive model for the forest fire sketched earlier, with endogenous variables L, ML, and F. We want to show that $L = 1$ is an actual cause of $F = 1$. Clearly $(M, (1, 1)) \models F = 1$ 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 $\vec{Z} = \{L, F\}$ and take $\vec{W} = \{ML\}$, let $x' = 0$, and let $w = 0$. Clearly, $(M, (1, 1)) \models [L \leftarrow 0, ML \leftarrow 0](F \neq 1)$; 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 ML to 0. (Note that here setting L and ML to 0 overrides the effects of U; this is critical.) Moreover, $(M, (1, 1)) \models [L \leftarrow 1, ML \leftarrow 0](F = 1)$; 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 $\vec{Z} = \{L, F\}$, the only subsets of $\vec{Z} - \vec{X}$ are the empty set and the singleton set consisting of just F.)

It is also straightforward to show that the lightning and the dropped match are also causes of the forest fire in the context where $U = (1, 1)$ in the conjunctive model. Again, AC1 and AC3 are trivially satisfied and, again, to show that AC2 holds in the case of lightning we can take $\vec{Z} = \{L, F\}$, $\vec{W} = \{ML\}$, and $x' = 0$, but now we let $w = 1$. In the conjunctive scenario, if there is no lightning, there is no forest fire, while if there is lightning (and the match is dropped) there is a forest fire, so AC2(a) and AC2(b) are satisfied; similarly for the dropped match.

EXAMPLE 3. Now consider the Suzy-Billy example.[4] We get the desired result—that Suzy's throw is a cause, but Billy's is not—but only if we model the story appropriately. Consider first a coarse causal model, with three endogenous variables:

- ST for "Suzy throws", with values 0 (Suzy does not throw) and 1 (she does);
- BT for "Billy throws", with values 0 (he doesn't) and 1 (he does);
- BS for "bottle shatters", with values 0 (it doesn't shatter) and 1 (it does).

(We omit the exogenous variable here; it determines whether Billy and Suzy throw.) Take the formula for BS to be such that the bottle shatters if either Billy or Suzy throw; that is $BS = \max(BT, ST)$. (We assume that Suzy and Billy will not miss if they throw.) BT and ST play symmetric roles in this model; there is nothing to distinguish them. Not surprisingly, both Billy's throw and Suzy's throw are classified as causes of the bottle shattering in this model. The argument is essentially identical to that in the disjunctive model of the forest-fire example in

[4]The discussion of this example is taken almost verbatim from HP.

the context $U = (1, 1)$, where both the lightning and the dropped match are causes of the fire.

The trouble with this model is that it cannot distinguish the case where both rocks hit the bottle simultaneously (in which case it would be reasonable to say that both $ST = 1$ and $BT = 1$ are causes of $BS = 1$) from the case where Suzy's rock hits first. To allow the model to express this distinction, we add two new variables to the model:

- BH for "Billy's rock hits the (intact) bottle", with values 0 (it doesn't) and 1 (it does); and
- SH for "Suzy's rock hits the bottle", again with values 0 and 1.

Now our equations will include:

- $SH = ST$;
- $BH = \min(BT, 1 - SH)$; and
- $BS = \max(SH, BH)$.

Now it is the case that, in the context where both Billy and Suzy throw, $ST = 1$ is a cause of $BS = 1$, but $BT = 1$ is not. To see that $ST = 1$ is a cause, note that, as usual, it is immediate that AC1 and AC3 hold. For AC2, choose $\vec{Z} = \{ST, SH, BH, BS\}$, $\vec{W} = \{BT\}$, and $w = 0$. When BT is set to 0, BS tracks ST: if Suzy throws, the bottle shatters and if she doesn't throw, the bottle does not shatter. To see that $BT = 1$ is *not* a cause of $BS = 1$, we must check that there is no partition $\vec{Z} \cup \vec{W}$ of the endogenous variables that satisfies AC2. Attempting the symmetric choice with $\vec{Z} = \{BT, BH, SH, BS\}$, $\vec{W} = \{ST\}$, and $w = 0$ violates AC2(b). To see this, take $\vec{Z}' = \{BH\}$. In the context where Suzy and Billy both throw, $BH = 0$. If BH is set to 0, the bottle does not shatter if Billy throws and Suzy does not. It is precisely because, in this context, Suzy's throw hits the bottle and Billy's does not that we declare Suzy's throw to be the cause of the bottle shattering. AC2(b) captures that intuition by allowing us to consider the contingency where $BH = 0$, despite the fact that Billy throws. We leave it to the reader to check that no other partition of the endogenous variables satisfies AC2 either.

This example emphasizes an important moral. If we want to argue in a case of preemption that $X = x$ is the cause of ϕ rather than $Y = y$, then there must be a random variable (BH in this case) that takes on different values depending on whether $X = x$ or $Y = y$ is the actual cause. If the model does not contain such a variable, then it will not be possible to determine which one is in fact the cause. This is certainly consistent with intuition and the way we present evidence. If we want to argue (say, in a court of law) that it was A's shot that killed C rather than B's, then we present evidence such as the bullet entering C from the left side (rather

than the right side, which is how it would have entered had B's shot been the lethal one). The side from which the shot entered is the relevant random variable in this case. Note that the random variable may involve temporal evidence (if Y's shot had been the lethal one, the death would have occurred a few seconds later), but it certainly does not have to.

4 The Choice of Variables

A modeler has considerable leeway in choosing which variables to include in a model. Nature does not provide a uniquely correct set of variables. Nonetheless, there are a number of considerations that guide variable selection. While these will not usually suffice to single out one choice of variables, they can provide a framework for the rational evaluation of models, including resources for motivating and defending certain choices of variables, and criticizing others.

The problem of choosing a set of variables for inclusion in a model has many dimensions. One set of issues concerns the question of how many variables to include in a model. If the modeler begins with a set of variables, how can she know whether she should add additional variables to the model? Given that it is always possible to add additional variables, is there a point at which the model contains "enough" variables? Is it ever possible for a model to have "too many" variables? Can the addition of further variables ever do positive harm to a model?

Another set of issues concerns the values of variables. Say that variable X' is a *refinement* of X if, for each value x in the range of X, there is some subset S of the range of X' such that $X = x$ just in case X' is in S. When is it appropriate or desirable to replace a variable with a refinement? Can it ever lead to problems if a variable is too fine-grained? Similarly, are there considerations that would lead us to prefer a model that replaced X with a new variable X'', whose range is a proper subset or superset of the range of X?

Finally, are there constraints on the set of variables in a model over and above those we might impose on individual variables? For instance, can the choice to include a particular variable X within a model require us to include another variable Y, or to exclude a particular variable Z?

While we cannot provide complete answers to all of these questions, we believe a good deal can be said to reduce the arbitrariness of the choice of variables. The most plausible way to motivate guidelines for the selection of variables is to show how inappropriate choices give rise to systems of equations that are inaccurate, misleading, or incomplete in their predictions of observations and interventions. In the next three subsections, we present several examples to show how such considerations can be brought to bear on the problem of variable choice.

4.1 The Number of Variables

We already saw in Example 3 that it is important to choose the variables correctly. Adding more variables can clearly affect whether A is a cause of B. When is it

appropriate or necessary to add further variables to a model?[5] Suppose that we have an infinite sequence of models $M^1, M^2, \ldots$ such that the variables in M^i are $X_0, \ldots, X_{i+1}, Y$, and $M^{i+1}_{X_{i+1} \leftarrow 1} = M_i$ (so that M^{i+1} can be viewed as an extension of M^i). Is it possible that whether $X_0 = 1$ is a cause of $Y = 1$ can alternate as we go through this sequence? This would indicate a certain "instability" in the causality. In this circumstance, a lawyer should certainly be able to argue against using, say, M^7 as a model to show that $X_0 = 1$ is cause of $Y = 1$. On the other hand, if the sequence stabilizes, that is, if there is some k such that for all $i \geq k$, M^i delivers the same verdict on some causal claim of interest, that would provide a strong reason to accept M^k as sufficient.

Compare Example 2 with Example 3. In Example 2, we were able to adequately model the scenario using only three endogenous variables: L, ML, and F. By contrast, in Example 3, the model containing only three endogenous variables, BT, ST, and BS, was inadequate. What is the difference between the two scenarios? One difference we have already mentioned is that there seems to be an important feature of the second scenario that cannot be captured in the three-variable model: Suzy's rock hit the bottle before Billy's did. There is also a significant "topological" difference between the two scenarios. In the forest-fire example, there are two directed paths into the variable F. We could interpolate additional variables along these two paths. We could, for instance, interpolate a variable representing the occurrence of a small brush fire. But doing so would not fundamentally change the causal structure: there would still be just two directed paths into F. In the case of preemption, however, adding the additional variables SH and BH created an additional directed path that was not there before. The three-variable model contained just two directed paths: one from ST to BS, and one from BT to BS. However, once the variables SH and BH were added, there were three directed paths: $\{ST, SH, BS\}$, $\{BT, BH, BS\}$, and $\{ST, SH, BH, BS\}$. The intuition, then, is that adding additional variables to a model will not affect the relations of actual causation that hold in the model unless the addition of those variables changes the "topology" of the model. A more complete mathematical characterization of the conditions under which the verdicts of actual causality remain stable under the addition of further variables strikes us as a worthwhile research project that has not yet been undertaken.

4.2 The Ranges of Variables

Not surprisingly, the set of possible values of a variable must also be chosen appropriately. Consider, for example, a case of "trumping", introduced by Schaffer [2000]. Suppose that a group of soldiers is very well trained, so that they will obey any order given by a superior officer; in the case of conflicting orders, they obey the

[5] Although his model of causality is quite different from ours, Spohn [2003] also considers the effect of adding or removing variables, and discusses how a model with fewer variables should be related to one with more variables.

highest-ranking officer. Both a sergeant and a major issue the order to march, and the soldiers march. Let us put aside the morals that Schaffer attempts to draw from this example (with which we disagree; see [Halpern and Pearl 2005] and [Hitchcock 2010]), and consider only the modeling problem. We will presumably want variables S, M, and A, corresponding to the sergeant's order, the major's order, and the soldiers' action. We might let $S = 1$ represent the sergeant's giving the order to march and $S = 0$ represent the sergeant's giving no order; likewise for M and A. But this would not be adequate. If the only possible order is the order to march, then there is no way to capture the principle that in the case of conflicting orders, the soldiers obey the major. One way to do this is to replace the variables M, S, and A by variables M', S' and A' that take on three possible values. Like M, $M' = 0$ if the major gives no order and $M' = 1$ if the major gives the order to march. But now we allow $M' = 2$, which corresponds to the major giving some other order. S' and A' are defined similarly. We can now write an equation to capture the fact that if $M' = 1$ and $S' = 2$, then the soldiers march, while if $M' = 2$ and $S' = 1$, then the soldiers do not march.

The appropriate set of values of a variable will depend on the other variables in the picture, and the relationship between them. Suppose, for example, that a hapless homeowner comes home from a trip to find that his front door is stuck. If he pushes on it with a normal force then the door will not open. However, if he leans his shoulder against it and gives a solid push, then the door will open. To model this, it suffices to have a variable O with values either 0 or 1, depending on whether the door opens, and a variable P, with values 0 or 1 depending on whether or not the homeowner gives a solid push.

On the other hand, suppose that the homeowner also forgot to disarm the security system, and that the system is very sensitive, so that it will be tripped by any push on the door, regardless of whether the door opens. Let $A = 1$ if the alarm goes off, $A = 0$ otherwise. Now if we try to model the situation with the same variable P, we will not be able to express the dependence of the alarm on the homeowner's push. To deal with both O and A, we need to extend P to a 3-valued variable P', with values 0 if the homeowner does not push the door, 1 if he pushes it with normal force, and 2 if he gives it a solid push.

These considerations parallel issues that arise in philosophical discussions about the metaphysics of "events".[6] Suppose that our homeowner pushed on the door with enough force to open it. Is there just one event, the push, that can be described at various levels of detail, such as a "push" or a "hard push"? This is the view of Davidson [1967]. Or are there rather many different events corresponding to these different descriptions, as argued by Kim [1973] and Lewis [1986b]? And if we take the latter view, which of the many events that occur should be counted as causes of the door's opening? These strike us as pseudoproblems. We believe that questions

[6]This philosophical usage of the word "event" is different from the typical usage of the word in computer science and probability, where an event is just a subset of the state space.

about causality are best addressed by dealing with the methodological problem of constructing a model that correctly describes the effects of interventions in a way that is not misleading or ambiguous.

A slightly different way in which one variable may constrain the values that another may take is by its implicit presuppositions. For example, a counterfactual theory of causation seems to have the somewhat counterintuitive consequence that one's birth is a cause of one's death. This sounds a little odd. If Jones dies suddenly one night, shortly before his 80th birthday, the coroner's inquest is unlikely to list "birth" as among the causes of his death. Typically, when we investigate the causes of death, we are interested in what makes the difference between a person's dying and his surviving. So our model might include a variable D such $D = 1$ holds if Jones dies shortly before his 80th birthday, and $D = 0$ holds if he continues to live. If our model also includes a variable B, taking the value 1 if Jones is born, 0 otherwise, then there simply is no value that D would take if $B = 0$. Both $D = 0$ and $D = 1$ implicitly presuppose that Jones was born (i.e., $B = 1$). Our conclusion is that if we have chosen to include a variable such as D in our model, then we cannot conclude that Jones' birth is a cause of his death!

4.3 Dependence and Independence

Lewis [1986a] added a constraint to his counterfactual theory of causation. In order for event c to be a cause of event e, the two events cannot be logically related. Suppose for instance, that Martha says "hello" loudly. If she had not said "hello", then she certainly could not have said "hello" loudly. But her saying "hello" is not a cause of her saying "hello" loudly. The counterfactual dependence results from a logical, rather than a causal, relationship between the two events.

We must impose a similar constraint upon causal models. Values of different variables should not correspond to events that are logically related. But now, rather than being an *ad hoc* restriction, it has a clear rationale. For suppose that we had a model with variable H_1 and H_2, where H_1 represents "Martha says 'hello'" (i.e., $H_1 = 1$ if Martha says "hello" and $H_1 = 0$ otherwise), and H_2 represents "Martha says 'hello' loudly". The intervention $H_1 = 0 \wedge H_2 = 1$ is meaningless; it is logically impossible for Martha not to say "hello" and to say ''hello" loudly.

We doubt that any careful modeler would choose variables that have logically related values. However, the converse of this principle, that the different values of any particular variable *should* be logically related (in fact, mutually exclusive), is less obvious and equally important. Consider Example 3. While, in the actual context, Billy's rock will hit the bottle just in case Suzy's doesn't, this is not a necessary relationship. Suppose that, instead of using two variables SH and BH, we try to model the scenario with a variable H that takes the value 1 if Suzy's rock hits, and and 0 if Billy's rock hits. The reader can verify that, in this model, there is no contingency such that the bottle's shattering depends upon Suzy's throw. The problem, as we said, is that $H = 0$ and $H = 1$ are *not* mutually exclusive; there are

possible situations in which both rocks hit or neither rock hits the bottle. In particular, this representation does not allow us to consider independent interventions on the rocks hitting the bottle. As the discussion in Example 3 shows, it is precisely such an intervention that is needed to establish that Suzy's throw (and not Billy's) is the actual cause of the bottle shattering.

While these rules are simple in principle, their application is not always transparent.

EXAMPLE 4. Consider cases of "switching", which have been much discussed in the philosophical literature. A train is heading toward the station. An engineer throws a switch, directing the train down the left track, rather than the right track. The tracks re-converge before the station, and the train arrives as scheduled. Was throwing the switch a cause of the train's arrival? HP consider two causal models of this scenario. In the first, there is a random variable S which is 1 if the switch is thrown (so the train goes down the left track) and 0 otherwise. In the second, in addition to S, there are variables LT and RT, indicating whether or not the train goes down the left track and right track, respectively. Note that with the first representation, there is no way to model the train not making it to the arrival point. With the second representation, we have the problem that $LT = 1$ and $RT = 1$ are arguably not independent; the train cannot be on both tracks at once. If we want to model the possibility of one track or another being blocked, we should use, instead of LT and RT, variables LB and RB, which indicate whether the left track or right track, respectively, are blocked. This allows us to represent all the relevant possibilities without running into independence problems. Note that if we have only S as a random variable, then $S = 1$ cannot be a cause of the train arriving; it would have arrived no matter what. With RB in the picture, the preliminary HP definition of actual cause rules that $S = 1$ can be an actual cause of the train's arrival; for example, under the contingency that $RB = 1$, the train does not arrive if $S = 0$. (However, once we extend the definition to include defaults, as we will in the next section, it becomes possible once again to block this conclusion.)

These rules will have particular consequences for how we should represent events that might occur at different times. Consider the following simplification of an example introduced by Bennett [1987], and also considered in HP.

EXAMPLE 5. Suppose that the Careless Camper (CC for short) has plans to go camping on the first weekend in June. He will go camping unless there is a fire in the forest in May. If he goes camping, he will leave a campfire unattended, and there will be a forest fire. Let the variable C take the value 1 if CC goes camping, and 0 otherwise. How should we represent the state of the forest?

There appear to be at least three alternatives. The simplest proposal would be to use a variable F that takes the value 1 if there is a forest fire at some time, and 0 otherwise.[7] But now how are we to represent the dependency relations between F

[7] This is, in effect, how effects have been represented using "neuron diagrams" in late preemption

and C? Since CC will go camping only if there is no fire (in May), we would want to have an equation such as $C = 1 - F$. On the other hand, since there will be a fire (in June) just in case CC goes camping, we will also need $F = C$. This representation is clearly not rich enough, since it does not let us make the clearly relevant distinction between whether the forest fire occurs in May or June. The problem is manifested in the fact that the equations are cyclic, and have no consistent solution.[8]

A second alternative, adopted by Halpern and Pearl [2005, p. 860], would be to use a variable F' that takes the value 0 if there is no fire, 1 if there is a fire in May, and 2 if there is a fire in June. Now how should we write our equations? Since CC will go camping unless there is a fire in May, the equation for C should say that $C = 0$ iff $F' = 1$. And since there will be a fire in June if CC goes camping, the equation for F' should say that $F' = 2$ if $C = 1$ and $F' = 0$ otherwise. These equations are cyclic. Moreover, while they do have a consistent solution, they are highly misleading in what they predict about the effects of interventions. For example, the first equation tells us that intervening to create a forest fire in June would cause CC to go camping in the beginning of June. But this seems to get the causal order backwards!

The third way to model the scenario is to use two separate variables, F_1 and F_2, to represent the state of the forest at separate times. $F_1 = 1$ will represent a fire in May, and $F_1 = 0$ represents no fire in May; $F_2 = 1$ represents a fire in June and $F_2 = 0$ represents no fire in June. Now we can write our equations as $C = 1 - F_1$ and $F_2 = C \times (1 - F_1)$. This representation is free from the defects that plague the other two representations. We have no cycles, and hence there will be a consistent solution for any value of the exogenous variables. Moreover, this model correctly tells us that only an intervention on the state of the forest in May will affect CC's camping plans.

Once again, our discussion of the methodology of modeling parallels certain metaphysical discussions in the philosophy literature. If heavy rains delay the onset of a fire, is it the same fire that would have occurred without the rains, or a different fire? It is hard to see how to gain traction on such an issue by direct metaphysical speculation. By contrast, when we recast the issue as one about what kinds of variables to include in causal models, it is possible to say exactly how the models will mislead you if you make the wrong choice.

---

cases. See Hitchcock [2007b, pp. 85–88] for discussion.

[8] Careful readers will note the the preemption case of Example 3 is modeled in this way. In that model, BH is a cause of BS, even though it is the earlier shattering of the bottle that prevents Billy's rock from hitting. Halpern and Pearl [2005] note this problem and offer a dynamic model akin to the one recommended below. As it turns out, this does not affect the analysis of the example offered above.

5 Dealing with normality and typicality

While the definition of causality given in Definition 1 works well in many cases, it does not always deliver answers that agree with (most people's) intuition. Consider the following example, taken from Hitchcock [2007a], based on an example due to Hiddleston [2005].

EXAMPLE 6. 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 preliminary HP definition, it is. For in the contingency where Assassin puts in the poison, Victim survives iff Bodyguard puts in the antidote.

Example 6 illustrates an even deeper problem with Definition 1. The structural equations for Example 6 are *isomorphic* to those in the forest-fire example, provided that we interpret the variables appropriately. Specifically, take the endogenous variables in Example 6 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, ML, and F, respectively. In the context where there is lightning and the arsonists drops a lit match, both 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. (An example illustrating the same phenomenon is given by Hall [2007].) This suggests that there must be more to causality than just the structural equations. And, indeed, the final HP definition of causality allows certain contingencies to be labeled as "unreasonable" or "too farfetched"; these contingencies are then not considered in AC2(a) or AC2(b). As discussed by Halpern [2008], there are problems with the HP account; we present here the approach used in [Halpern 2008] for dealing with these problems, which involves assuming that an agent has, in addition to a theory of causality (as modeled by the structural equations), a theory of "normality" or "typicality". (The need to consider normality was also stressed by Hitchcock [2007a] and Hall [2007], and further explored by Hitchcock and Knobe [2009].) This theory would include statements like "typically, people do not put poison in coffee" and "typically doctors do not treat patients to whom they are not assigned". There are many ways of giving semantics to such typicality statements (e.g., [Adams 1975; Kraus, Lehmann, and Magidor 1990; Spohn 2009]). For definiteness, we use *ranking functions* [Spohn 2009] here.

Take a *world* to be a complete description of the values of all the random variables. we assume that each world has associated with it a *rank*, which is just a natural number or ∞. Intuitively, the higher the rank, the less "normal" or "typical" the

world. A world with a rank of 0 is reasonably normal, one with a rank of 1 is somewhat normal, one with a rank of 2 is quite abnormal, and so on. Given a ranking on worlds, the statement "if p then typically q" is true if in all the worlds of least rank where p is true, q is also true. Thus, in one model where people do not typically put either poison or antidote in coffee, the worlds where neither poison nor antidote is put in the coffee have rank 0, worlds where either poison or antidote is put in the coffee have rank 1, and worlds where both poison and antidote are put in the coffee have rank 2.

Take an *extended causal model* to be a tuple $M = (\mathcal{S}, \mathcal{F}, \kappa)$, where $(\mathcal{S}, \mathcal{F})$ is a causal model, and κ is a *ranking function* that associates with each world a rank. In an acyclic extended causal model, a context $\vec{u}$ determines a world, denoted $s_{\vec{u}}$. $\vec{X} = \vec{x}$ is 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 1, except that in AC2(a), there must be a world s such that $\kappa(s) \leq \kappa(s_{\vec{u}})$ and $\vec{X} = \vec{x}' \wedge \vec{W} = \vec{w}$ is true at s. This can be viewed as a formalization of Kahneman and Miller's [1986] observation that "an event is more likely to be undone by altering exceptional than routine aspects of the causal chain that led to it".

This definition deals well with all the problematic examples in the literature. Consider Example 6. Using the ranking described above, Bodyguard is not a cause of Victim's survival because the world that would need to be considered in AC2(a), where Assassin poisons the coffee, is less normal than the actual world, where he does not. We consider just one other example here (see [Halpern 2008] for further discussion).

EXAMPLE 7. Consider the following story, taken from (an early version of) [Hall 2004]: Suppose that Billy is hospitalized with a mild illness on Monday; he is treated and recovers. In the obvious causal model, the doctor's treatment is a cause of Billy's recovery. Moreover, if the doctor does *not* treat Billy on Monday, then the doctor's omission to treat Billy is a cause of Billy's being sick on Tuesday. But now suppose that there are 100 doctors in the hospital. Although only doctor 1 is assigned to Billy (and he forgot to give medication), in principle, any of the other 99 doctors could have given Billy his medication. Is the nontreatment by doctors 2–100 also a cause of Billy's being sick on Tuesday?

Suppose that in fact the hospital has 100 doctors and there are variables $A_1, \ldots, A_{100}$ and $T_1, \ldots, T_{100}$ in the causal model, where $A_i = 1$ if doctor i is assigned to treat Billy and $A_i = 0$ if he is not, and $T_i = 1$ if doctor i actually treats Billy on Monday, and $T_i = 0$ if he does not. Doctor 1 is assigned to treat Billy; the others are not. However, in fact, no doctor treats Billy. Further assume that, typically, no doctor is assigned to a given patient; if doctor i is not assigned to treat Billy, then typically doctor i does not treat Billy; and if doctor i *is* assigned to Billy, then typically doctor i treats Billy. We can capture this in an extended causal model where the world where no doctor is assigned to Billy and no doctor

treats him has rank 0; the 100 worlds where exactly one doctor is assigned to Billy, and that doctor treats him, have rank 1; the 100 worlds where exactly one doctor is assigned to Billy and no one treats him have rank 2; and the 100×99 worlds where exactly one doctor is assigned to Billy but some other doctor treats him have rank 3. (The ranking given to other worlds is irrelevant.) In this extended model, in the context where doctor i is assigned to Billy but no one treats him, i is the cause of Billy's sickness (the world where i treats Billy has lower rank than the world where i is assigned to Billy but no one treats him), but no other doctor is a cause of Billy's sickness. Moreover, in the context where i is assigned to Billy and treats him, then i is the cause of Billy's recovery (for AC2(a), consider the world where no doctor is assigned to Billy and none treat him).

Adding a normality theory to the model gives the HP account of actual causation greater flexibility to deal with these kinds of cases. This raises the worry, however, that this gives the modeler too much flexibility. After all, the modeler can now render any claim that A is an actual cause of B false, simply by choosing a normality order that assigns the actual world $s_{\vec{u}}$ a lower rank than any world s needed to satisfy AC2. Thus, the introduction of normality exacerbates the problem of motivating and defending a particular choice of model. Fortunately, the literature on the psychology of counterfactual reasoning and causal judgment goes some way toward enumerating the sorts of factors that constitute normality. (See, for example, [Alicke 1992; Cushman 2009; Cushman, Knobe, and Sinnott-Armstrong 2008; Hitchcock and Knobe 2009; Kahneman and Miller 1986; Knobe and Fraser 2008; Kahneman and Tversky 1982; Mandel, Hilton, and Catellani 1985; Roese 1997].) These factors include the following:

- Statistical norms concern what happens most often, or with the greatest frequency. Kahneman and Tversky [1982] gave subjects a story in which Mr. Jones usually leaves work at 5:30, but occasionally leaves early to run errands. Thus, a 5:30 departure is (statistically) "normal", and an earlier departure "abnormal". This difference affected which alternate possibilities subjects were willing to consider when reflecting on the causes of an accident in which Mr. Jones was involved.

- Norms can involve moral judgments. Cushman, Knobe, and Sinnott-Armstrong [2008] showed that people with different views about the morality of abortion have different views about the abnormality of insufficient care for a fetus, and this can lead them to make different judgments about the cause of a miscarriage.

- Policies adopted by social institutions can also be norms. For instance, Knobe and Fraser [2008] presented subjects with a hypothetical situation in which a department had implemented a policy allowing administrative assistants to take pens from the department office, but prohibiting faculty from doing

so. Subjects were more likely to attribute causality to a professor's taking a pen than to an assistant's taking one, even when the situation was otherwise similar.

- There can also be norms of "proper functioning" governing the operations of biological organs or mechanical parts: there are certain ways that hearts and spark plugs are "supposed" to operate. Hitchcock and Knobe [2009] show that these kinds of norms can also affect causal judgments.

The law suggests a variety of principles for determining the norms that are used in the evaluation of actual causation. In criminal law, norms are determined by direct legislation. For example, if there are legal standards for the strength of seat belts in an automobile, a seat belt that did not meet this standard could be judged a cause of a traffic fatality. By contrast, if a seat belt complied with the legal standard, but nonetheless broke because of the extreme forces it was subjected to during a particular accident, the fatality would be blamed on the circumstances of the accident, rather than the seat belt. In such a case, the manufacturers of the seat belt would not be guilty of criminal negligence. In contract law, compliance with the terms of a contract has the force of a norm. In tort law, actions are often judged against the standard of "the reasonable person". For instance, if a bystander was harmed when a pedestrian who was legally crossing the street suddenly jumped out of the way of an oncoming car, the pedestrian would not be held liable for damages to the bystander, since he acted as the hypothetical "reasonable person" would have done in similar circumstances. (See, for example, [Hart and Honoré 1985, pp. 142ff.] for discussion.) There are also a number of circumstances in which deliberate malicious acts of third parties are considered to be "abnormal" interventions, and affect the assessment of causation. (See, for example, [Hart and Honoré 1985, pp. 68ff.].)

As with the choice of variables, we do not expect that these considerations will always suffice to pick out a uniquely correct theory of normality for a causal model. They do, however, provide resources for a rational critique of models.

6 Conclusion

As HP stress, causality is relative to a model. That makes it particularly important to justify whatever model is chosen, and to enunciate principles for what makes a reasonable causal model. We have taken some preliminary steps in investigating this issue with regard to the choice of variables and the choice of defaults. However, we hope that we have convinced the reader that far more needs to be done if causal models are actually going to be used in applications.

Acknowledgments: We thank Wolfgang Spohn for useful comments. Joseph Halpern was supported in part by NSF grants IIS-0534064 and IIS-0812045, and by AFOSR grants FA9550-08-1-0438 and FA9550-05-1-0055.

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.

Bennett, J. (1987). Event causation: the counterfactual analysis. In *Philosophical Perspectives, Vol. 1, Metaphysics*, pp. 367–386. Atascadero, CA: Ridgeview Publishing Company.

Cushman, F. (2009). The role of moral judgment in causal and intentional attribution: What we say or how we think?". Unpublished manuscript.

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

Davidson, D. (1967). Causal relations. *Journal of Philosophy LXIV*(21), 691–703.

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.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica 40*(6), 979–1001.

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

Hansson, R. N. (1958). *Patterns of Discovery*. Cambridge, U.K.: Cambridge University Press.

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. (2007a). Prevention, preemption, and the principle of sufficient reason. *Philosophical Review 116*, 495–532.

Hitchcock, C. (2007b). What's wrong with neuron diagrams? In J. Campbell, M. O'Rourke, and H. Silverstein (Eds.), *Causation and Explanation*, pp. 69–92. Cambridge, MA: MIT Press.

Hitchcock, C. (2010). Trumping and contrastive causation. *Synthese*. To appear.

Hitchcock, C. and J. Knobe (2009). Cause and norm. *Journal of Philosophy*. To appear.

Hume, D. (1739). *A Treatise of Human Nature*. London: John Noon.

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.

Kim, J. (1973). Causes, nomic subsumption, and the concept of event. *Journal of Philosophy LXX*, 217–236.

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. (1973a). Causation. *Journal of Philosophy 70*, 113–126. Reprinted with added "Postscripts" in D. Lewis, *Philosophical Papers*, Volume II, Oxford University Press, 1986, pp. 159–213.

Lewis, D. (1986a). 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. (1986b). Events. In *Philosophical Papers*, Volume II, pp. 241–270. New York: Oxford University Press.

Lewis, D. K. (1973b). *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

Mandel, D. R., D. J. Hilton, and P. Catellani (Eds.) (1985). *The Psychology of Counterfactual Thinking*. New York: Routledge.

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

Roese, N. (1997). Counterfactual thinking. *Psychological Bulletin CXXI*, 133–148.

Schaffer, J. (2000). Trumping preemption. *Journal of Philosophy XCVII*(4), 165–181. Reprinted in J. Collins and N. Hall and L. A. Paul (eds.), *Causation and Counterfactuals*, MIT Press, 2002.

Spohn, W. (2003). Dependency equilibria and the causal structure of decision and game situations. In *Homo Oeconomicus XX*, pp. 195–255.

Spohn, W. (2009). A survey of ranking theory. In F. Huber and C. Schmidt-Petri (Eds.), *Degrees of Belief. An Anthology*, pp. 185–228. Dordrecht, Netherlands: Springer.

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