Causes and Explanations: A Structural-Model Approach. Part I: Causes

Joseph Y. Halpern*
Cornell University
Dept. of Computer Science
Ithaca, NY 14853
halpern@cs.cornell.edu
http://www.cs.cornell.edu/home/halpern

Judea Pearl†
Dept. of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
judea@cs.ucla.edu
http://www.cs.ucla.edu/∼judea

October 24, 2005

Abstract

We propose a new definition of *actual causes*, using *structural equations* to model counterfactuals. We show that the definition yields a plausible and elegant account of causation that handles well examples which have caused problems for other definitions and resolves major difficulties in the traditional account.

*Supported in part by NSF under grant IRI-96-25901
†Supported in part by grants from NSF, ONR, AFOSR, and MICRO.

1 Introduction

What does it mean that an event C *actually caused* event E? The problem of defining "actual cause" goes beyond mere philosophical speculation. As Good [1993] and Michie [1999] argue persuasively, in many legal settings, what needs to be established (for determining responsibility) is exactly such "cause in fact". A typical example [Wright 1988] considers two fires advancing toward a house. If fire A burned the house before fire B, we (and many juries nationwide) would consider fire A "the actual cause" for the damage, even supposing that the house would definitely have been burned down by fire B, if it were not for A. Actual causation is also important in artificial intelligence applications. Whenever we undertake to *explain* a set of events that unfold in a specific scenario, the explanation produced must acknowledge the actual cause of those events. The automatic generation of adequate explanations, a task essential in planning, diagnosis, and natural language processing, therefore requires a formal analysis of the concept of actual cause.

The philosophy literature has been struggling with this problem of defining causality since at least the days of Hume [1739], who was the first to identify causation with counterfactual dependence. To quote Hume [1748, Section VIII]:

> 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.

Among modern philosophers, the counterfactual interpretation of causality continues to receive most attention, primarily due to the work of David Lewis [1973]. Lewis has given counterfactual dependence formal underpinning in possible-world semantics and has equated actual causation with the transitive closure of counterfactual dependencies. C is classified as a cause of E if C is linked to E by a chain of events each directly depending on its predecessor. However, Lewis's dependence theory has encountered many difficulties. (See [Collins, Hall, and Paul 2004; Hall and Paul 2003; Pearl 2000; Sosa and Tooley 1993] for some recent discussion.) The problem is that effects may not always counterfactually depend on their causes, either directly or indirectly, as the two-fire example illustrates. In addition, causation is not always transitive, as implied Lewis's chain-dependence account (see Example 4.3).

Here we give a definition of actual causality cast in the language of *structural equations*. The basic idea is to extend the basic notion of counterfactual dependency to allow "contingent dependency". In other words, while effects may not always counterfactually depend on their causes in the actual situation, they do depend on them under certain contingencies. In the case of the two fires, for example, the house burning down does depend on fire A under the contingency that firefighters reach the house any time between the actual arrival of fire A and that of fire B. Under that contingency, if fire A had not been started, the house would not have burned down. The house burning down also depends on fire A under the contingency that fire B was not started. But this leads to an obvious concern: the house burning down also depends on fire B under the contingency that fire A was not started. We do not want to consider this latter contingency. Roughly speaking, we want to allow only contingencies that do not interfere with active causal processes. Our formal definition of actual causality tries to make this precise.

In Part II of the paper [Halpern and Pearl 2004], we give a definition of *(causal) explanation* using the definition of causality. An explanation adds information to an agent's knowledge; very roughly, an explanation of φ is a minimal elaboration of events that suffice to cause φ even in the face of uncertainty about the actual situation.

The use of structural equations as a model for causal relationships is standard in the social sciences, and seems to go back to the work of Sewall Wright in the 1920s (see [Goldberger 1972] for a discussion); the particular framework that we use here is due to Pearl [1995], and is further developed in [Galles and Pearl 1997; Halpern 2000; Pearl 2000]. While it is hard to argue that our definition (or any other definition, for that matter) is the "right" definition, we show that it deals well with the difficulties that have plagued other approaches in the past, especially those exemplified by the rather extensive compendium of Hall and Paul [2003].

According to our definition, the truth of every claim must be evaluated relative to a particular model of the world; that is, our definition allows us to claim only that C causes E in a (particular context in a) particular structural model. It is possible to construct two closely related structural models such that C causes E in one and C does not cause E in the other. Among other things, the modeler must decide which variables (events) to reason about and which to leave in the background. We view this as a feature of our model, not a bug. It moves the question of actual causality to the right arena—debating which of two (or more) models of the world is a better representation of those aspects of the world that one wishes to capture and reason about. This, indeed, is the type of debate that goes on in informal (and legal) arguments all the time.

There has been extensive discussion about causality in the philosophy literature. To keep this paper to manageable length, we spend only minimal time describing other approaches and comparing ours with them. We refer the reader to [Hall and Paul 2003; Pearl 2000; Sosa and Tooley 1993; Spirtes, Glymour, and Scheines 1993] for details and criticism of the probabilistic and logical approaches to causality in the philosophy literature. (We do try to point out where our definition does better than perhaps the best-known approach, due to Lewis [1973, 2000], as well as some other recent approaches [Hall 2000; Paul 1998; Yablo 2002], in the course of discussing the examples.)

There has also been work in the AI literature on causality. Perhaps the closest to this are papers by Pearl and his colleagues that use the structural-model approach. The definition of causality in this paper was inspired by an earlier paper of Pearl's [1998] that defined actual causality in terms of a construction called a *causal beam*. The definition was later modified somewhat (see [Pearl 2000, Chapter 10]). The modifications were in fact largely due to the considerations addressed in this paper. The definition given here is more transparent and handles a number of cases better (see Example A.3 in the appendix).

Tian and Pearl [2000] give results on estimating (from empirical data) the probability that C is a *necessary* cause of E—that is, the probability that E would not have occurred if C had not occurred. Necessary causality is related to but different from actual causality, as the definitions should make clear. Other work (for example, [Heckerman and Shachter 1995]) focuses on when a random variable X is the cause of a random variable Y; by way of contrast, we focus on when an *event* such as $X = x$ causes an event such as $Y = y$. Considering when a random variable is the cause of another is perhaps more appropriate as a *prospective* notion of causality: could X potentially be a cause of changes in Y. Our notion is more appropriate for a *retrospective* notion of causality: given all the information relevant to a given scenario, was $X = x$ the actual cause of $Y = y$ in that scenario? Many of the subtleties that arise when dealing with events simply disappear if we look at causality at the level of random variables. Finally, there is also a great deal of work in AI on formal action theory (see, for example, [Lin 1995; Sandewall 1994; Reiter 2001]), which is concerned with the proper way of incorporating causal relationships into a knowledge base so as to guide actions. The focus of our work is quite different; we are concerned with extracting the actual causality relation from such a knowledge base, coupled with a specific scenario.

The best ways to judge the adequacy of an approach are the intuitive appeal of the definitions and how well it deals with examples; we believe that this paper shows that our approach fares well on both counts.

The remainder of the paper is organized as follows. In the next section, we review structural models. In Section 3 we give a preliminary definition of actual causality and show in Section 4 how it deals with some examples of causality that have been problematic for other accounts. We refine the definition slightly in Section 5, and show how the refinement handles further examples. We conclude in Section 6 with some discussion.

2 Causal Models: A Review

In this section we review the basic definitions of causal models, as defined in terms of structural equations, and the syntax and semantics of a language for reasoning about causality. We also briefly compare our approach with the more standard approaches to modeling causality used in the literature.

Causal Models: The description of causal models given here is taken from [Halpern 2000]; the reader is referred to [Galles and Pearl 1997; Halpern 2000; Pearl 2000] for more details, motivation, and intuition.

The basic picture here is that we are interested in the values of random variables. If X is a random variable, a typical event has the form $X = x$. (In terms of possible worlds, this just represents the set of possible worlds where X takes on value x, although the model does not describe the set of possible worlds.) Some random variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. Each equation represents a distinct mechanism (or law) in the world, one that may be modified (by external actions) without altering the others. In practice, it seems 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. It is these endogenous variables whose values are described by the structural equations.

Formally, 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*). In most of this paper (except the appendix) we assume that $\mathcal{V}$ is finite. A *causal model* (or *structural model*) over signature $\mathcal{S}$ is a tuple $M = (\mathcal{S}, \mathcal{F})$, where $\mathcal{F}$ associates with each variable $X \in \mathcal{V}$ a function denoted F_X such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$. F_X determines the value of X given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. For example, if $F_X(Y, Z, U) = Y + U$ (which we usually write as $X = Y + U$), then if $Y = 3$ and $U = 2$, then $X = 5$, regardless of how Z is set. These equations can be thought of as representing processes (or mechanisms) by which values are assigned to variables. Hence, like physical laws, they support a counterfactual interpretation. For example, the equation above claims that, in the context $U = u$, if Y were 4, then X would be $u + 4$ (which we write as $(M, u) \models [Y \leftarrow 4](X = u + 4)$), regardless of what values X, Y, and Z actually take in the real world.

The function $\mathcal{F}$ defines a set of *(modifiable) structural equations*, relating the values of the variables. Because F_X is a function, there is a unique value of X once we have set all the other variables. Notice that we have such functions only for the endogenous variables. The exogenous variables are taken as given; it is their effect on the endogenous variables (and the effect of the endogenous variables on each other) that we are modeling with the structural equations.

The counterfactual interpretation and the causal asymmetry associated with the structural equations are best seen when we consider external interventions (or spontaneous changes), under which some equations in F are modified. An equation such as $x = F_X(\vec{u}, y)$ should be thought of as saying that in a context where the exogenous variables have values $\vec{u}$, if Y were set to y by some means (not specified in the model), then X would take on the value x, as dictated by F_X. The same does not hold when we intervene directly on X; such an intervention amounts to assigning a value to X by external means, thus overruling the assignment specified by F_X. In this case, Y is no longer committed to tracking X according to F_X. Variables on the left-hand side of equations are treated differently from ones on the right-hand side.

For those more comfortable with thinking of counterfactuals in terms of possible worlds, this modification of equations may be given a simple "closest world" interpretation: the solution of the equations obtained by replacing the equation for Y with the equation $Y = y$, while leaving all other equations unaltered, gives the closest "world" to the actual world where $Y = y$. In this possible-world interpretation, the asymmetry embodied in the model says that if $X = x$ in the closest world to w where $Y = y$, it does not follow that $Y = y$ in the closest worlds to w where $X = x$. In terms of structural equations, this just says that if $X = x$ is the solution for X under the intervention $Y = y$, it does not follow that $Y = y$ is the solution for Y under the intervention $X = x$. Each of two interventions modifies the system of equations in a distinct way; the former modifies the equation in which Y stands on the left, while the latter modifies the equation in which X stands on the left.

In summary, the equal sign in a structural equation differs from algebraic equality; in addition to describing an equality relationship between variables, it acts as an assignment statement in programming languages, since it specifies the way variables' values are determined. This should become clearer in our examples.

Example 2.1: Suppose that we want to reason about a forest fire that could be caused by either lightning or a match lit by an arsonist. Then the causal model would have the following endogenous variables (and perhaps others):

- F for fire ($F = 1$ if there is one, $F = 0$ otherwise);
- L for lightning ($L = 1$ if lightning occurred, $L = 0$ otherwise);
- *ML* for match lit (*ML* $= 1$ if the match was lit, *ML* $= 0$ otherwise).

The set $\mathcal{U}$ of exogenous variables includes conditions that suffice to render all relationships deterministic (such as whether the wood is dry, whether there is enough oxygen in the air for the match to light, etc.). Suppose that $\vec{u}$ is a setting of the exogenous variables that makes a forest fire possible (i.e., the wood is sufficiently dry, there is oxygen in the air, and so on). Then, for example, $F_F(\vec{u}, L, \mathit{ML})$ is such that $F = 1$ if either $L = 1$ or $\mathit{ML} = 1$. Note that although the value of F depends on the values of L and *ML*, the value of L does not depend on the values of F and *ML*. ∎

As we said, a causal model has the resources to determine counterfactual effects. Given a causal model $M = (\mathcal{S}, \mathcal{F})$, a (possibly empty) vector $\vec{X}$ of variables in $\mathcal{V}$, and vectors $\vec{x}$ and $\vec{u}$ of values for the variables in $\vec{X}$ and $\mathcal{U}$, respectively, we can define a new causal model denoted $M_{\vec{X} \leftarrow \vec{x}}$ over the signature $\mathcal{S}_{\vec{X}} = (\mathcal{U}, \mathcal{V} - \vec{X}, \mathcal{R}|_{\mathcal{V}-\vec{X}})$.[1] $M_{\vec{X} \leftarrow \vec{x}}$ is called a *submodel* of M by Pearl [2000]. Intuitively, this is the causal model that results when the variables in $\vec{X}$ are set to $\vec{x}$ by by some external action that affects only the variables in $\vec{X}$; we do not model the action or its causes explicitly. Formally, $M_{\vec{X} \leftarrow \vec{x}} = (\mathcal{S}_{\vec{X}}, \mathcal{F}^{\vec{X} \leftarrow \vec{x}})$, where $F_Y^{\vec{X} \leftarrow \vec{x}}$ is obtained from F_Y by setting the values of the variables in $\vec{X}$ to $\vec{x}$. For example, if M is the structural model describing Example 2.1, then the model $M_{L \leftarrow 0}$ has the equation $F = \mathit{ML}$. The equation for F in $M_{L \leftarrow 0}$ no longer involves L; rather, it is determined by setting L to 0 in the equation for F in M. Moreover, there is no equation for L in $M_{L \leftarrow 0}$.

It may seem strange that we are trying to understand causality using causal models, which clearly already encode causal relationships. Our reasoning is not circular. Our aim is not to reduce causation to noncausal concepts, but to interpret questions about causes of specific events in fully specified scenarios in terms of generic causal knowledge such as what we obtain from the equations of physics. The causal models encode background knowledge about the tendency of certain event types to cause other event types (such as the fact that lightning can cause forest fires). We use the models to determine the causes of single (or token) events, such

[1] We are implicitly identifying the vector $\vec{X}$ with the subset of $\mathcal{V}$ consisting of the variables in $\vec{X}$. $\mathcal{R}|_{\mathcal{V}-\vec{X}}$ is the restriction of $\mathcal{R}$ to the variables in $\mathcal{V} - \vec{X}$.

as whether it was arson that caused the fire of June 10, 2000, given what is known or assumed about that particular fire.

Notice that, in general, there may not be a unique vector of values that simultaneously satisfies the equations in $M_{\vec{X}\leftarrow\vec{x}}$; indeed, there may not be a solution at all. For simplicity in this paper, we restrict attention to what are called *recursive* (or *acyclic*) equations. This is the special case where there is some total ordering $\prec$ of the variables in $\mathcal{V}$ such that if $X \prec Y$, then F_X is independent of the value of Y; that is, $F_X(\ldots, y, \ldots) = F_X(\ldots, y', \ldots)$ for all $y, y' \in \mathcal{R}(Y)$. Intuitively, if a theory is recursive, there is no feedback. If $X \prec Y$, then the value of X may effect the value of Y, but the value of Y has no effect on the value of X. We do not lose much generality by restricting to recursive models (that is, ones whose equations are recursive). As suggested in the latter half of Example 4.2, it is always possible to timestamp events to impose an ordering on variables and thus construct a recursive model corresponding to a story. In any case, in the appendix, we sketch the necessary modifications of our definitions to deal with nonrecursive models.

It should be clear that if M is a recursive causal model, then there is always a unique solution to the equations in $M_{\vec{X}\leftarrow\vec{x}}$, given a setting $\vec{u}$ for the variables in $\mathcal{U}$ (we call such a setting $\vec{u}$ a *context*). We simply solve for the variables in the order given by $\prec$.

We can describe (some salient features of) a causal model M using a *causal network*. This is a graph with nodes corresponding to the random variables in $\mathcal{V}$ and an edge from a node labeled X to one labeled Y if F_Y depends on the value of X. This graph is a *dag*—a directed, acyclic graph (that is, a graph with no cycle of directed edges). The acyclicity follows from the assumption that the equations are recursive. Intuitively, variables can have a causal effect only on their descendants in the causal network; if Y is not a descendant of X, then a change in the value of X has no affect on the value of Y. For example, the causal network for Example 2.1 has the following form:

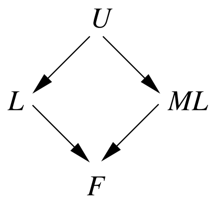

Figure 1: A simple causal network.

We remark that we occasionally omit the exogenous variables $\vec{U}$ from the causal network.

These causal networks, which are similar in spirit to the Bayesian networks used to represent and reason about dependences in probability distributions [Pearl 1988], will play a significant role in our definitions. They are quite similar in spirit to Lewis's [1973] *neuron diagrams*, but there are significant differences as well. Roughly speaking, neuron diagrams display explicitly the functional relationships (among variables in $\mathcal{V}$) for each specific context $\vec{u}$. The class of functions represented by neuron diagrams is limited to those described by "stimulatory" and "inhibitory" binary inputs. Causal networks represent arbitrary functional relationships, although the exact nature of the functions is specified in the structural equations and is not encoded in the diagram. The structural equations carry all the information we need to do causal reasoning, including all the information about belief, causation, intervention, and counterfactual behavior.

As we shall see, there are many nontrivial decisions to be made when choosing the structural model. 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. By way of contrast, in the philosophy literature, these events can be created on the fly, as it were. We return to this point in our examples.

Once the set of variables is chosen, it must be decided which are exogenous and which are endogenous. The exogenous variables to some extent encode the background situation, which we wish 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 wood is wet) and 1 (the wood is dry).[2] By making D exogenous, its value is assumed to be given and out of the control of the modeler. 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 oxygen explicitly at all. For example, suppose we have, as before, a random variable *ML* for match lit, and another variable *WB* for wood burning, with values 0 (it's not) and 1 (it is). The structural equation F_{WB} would describe the dependence of *WB* on D and *ML*. By setting $F_{WB}(1, 1) = 1$, we are saying that the wood will burn if the match is lit and the wood is dry. Thus, the equation is implicitly modeling our assumption that there is sufficient oxygen for the wood to burn.

We remark that, according to the definition in Section 3, only endogenous variables can be causes or be caused. Thus, if no variables encode the presence of oxygen, or if it is encoded only in an exogenous variable, then oxygen cannot be a cause of the wood burning. If we were to explicitly model the amount of oxygen in the air (which certainly might be relevant if we were analyzing fires on Mount Everest), then F_{WB} would also take values of O as an argument, and the presence of sufficient oxygen might well be a cause of the wood burning.[3]

Besides encoding some of our implicit assumptions, the structural equations can be viewed as encoding the causal mechanisms at work. Changing the underlying causal mechanism can affect what counts as a cause. Section 4 provides several examples of the importance of the choice of random variables and the choice of causal mechanism. It is not always straightforward to decide what the "right" causal model is in a given situation, nor is it always obvious which of two causal models is "better" in some sense. These may be difficult decisions and

[2]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.

[3]If there are other variables in the model, these would be arguments to F_{WB} as well; we have ignored other variables here just to make our point.

often lie at the heart of determining actual causality in the real world. Nevertheless, we believe that the tools we provide here should help in making principled decisions about those choices.

Syntax and Semantics: To make the definition of actual causality precise, it is helpful to have a logic with a formal syntax. Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a formula of the form $X = x$, for $X \in V$ and $x \in \mathcal{R}(X)$, is called a *primitive event*. A *basic causal formula (over $\mathcal{S}$)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$, where

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

Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\varphi$. The special case where $k = 0$ is abbreviated as φ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$ says that φ holds in the counterfactual world that would arise if Y_i were set to y_i, $i = 1, \ldots, k$. A *causal formula* is a Boolean combination of basic causal formulas.[4]

A causal formula ψ is true or false in a causal model, given a context. We write $(M, \vec{u}) \models \psi$ if ψ is true in causal model M given context $\vec{u}$.[5] $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}](X = x)$ if the variable X has value x in the unique (since we are dealing with recursive models) solution to the equations in $M_{\vec{Y} \leftarrow \vec{y}}$ in context $\vec{u}$ (that is, the unique vector of values for the exogenous variables that simultaneously satisfies all equations $F_Z^{\vec{Y} \leftarrow \vec{y}}$, $Z \in \mathcal{V} - \vec{Y}$, with the variables in $\mathcal{U}$ set to $\vec{u}$). $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ for an arbitrary Boolean combination φ of formulas of the form $\vec{X} = \vec{x}$ is defined similarly. We extend the definition to arbitrary causal formulas, i.e., Boolean combinations of basic causal formulas, in the obvious way.

Note that the structural equations are deterministic. We can make sense out of probabilistic counterfactual statements, even conditional ones (the probability that X would be 3 if Y_1 were 2, given that Y is in fact 1) in this framework (see [Balke and Pearl 1994]), by putting a probability on the set of possible contexts. This will not be necessary for our discussion of causality, although it will play a more significant role in the discussion of explanation.

3 The Definition of Cause

With all this notation in hand, we can now give a preliminary version of the definition of actual cause ("cause" for short). We want to make sense out of statements of the form "event A is

[4] If we write $\rightarrow$ for conditional implication, then a formula such as $[Y \leftarrow y]\varphi$ can be written as $Y = y \rightarrow \varphi$: if Y were y, then φ would hold. We use the present notation to emphasize the fact that, although we are viewing $Y \leftarrow y$ as a modal operator, we are not giving semantics using the standard possible worlds approach.

[5] We remark that in [Galles and Pearl 1997; Halpern 2000], the context $\vec{u}$ does not appear on the left-hand side of $\models$; rather, it is incorporated in the formula ψ on the right-hand side (so that a basic formula becomes $X(\vec{u}) = x$). Additionally, Pearl [2000] abbreviated $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}](X = x)$ as $X_y(u) = x$. The presentation here makes certain things more explicit, although they are technically equivalent.

an actual cause of event φ (in context $\vec{u}$)".[6] As we said earlier, the context is the background information. While this has been left implicit in some treatments of causality, we find it useful to make it explicit. The picture here is that the context (and the structural equations) are given. Intuitively, they encode the background knowledge. All the relevant events are known. The only question is picking out which of them are the causes of φ or, alternatively, testing whether a given set of events can be considered the cause of φ.[7]

The types of events that we allow 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; we typically abbreviate this as $\vec{X} = \vec{x}$. The events that can be caused are arbitrary Boolean combinations of primitive events. We might consider generalizing further to allow disjunctive causes. We do not believe that we lose much by disallowing disjunctive causes here. Since for causality we are assuming that the structural model and all the relevant facts are known, the only reasonable definition of "A or B causes φ" seems to be that "either A causes φ or B causes φ". There are no truly disjunctive causes once all the relevant facts are known.[8]

Definition 3.1: (Actual cause; preliminary version) $\vec{X} = \vec{x}$ is an *actual cause of* φ *in* $(M, \vec{u})$ if the following three conditions hold.

AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \varphi$. (That is, both $\vec{X} = \vec{x}$ and φ are true in the actual world.)

AC2. There exists a partition $(\vec{Z}, \vec{W})$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w}')$ of the variables in $(\vec{X}, \vec{W})$ such that if $(M, \vec{u}) \models Z = z^*$ for all $Z \in \vec{Z}$, then both of the following conditions hold:

(a) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}']\neg\varphi$. In words, changing $(\vec{X}, \vec{W})$ from $(\vec{x}, \vec{w})$ to $(\vec{x}', \vec{w}')$ changes φ from true to false.

(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W'} \leftarrow \vec{w}', \vec{Z'} \leftarrow \vec{z}^*]\varphi$ for all subsets $\vec{W'}$ of $\vec{W}$ and all subsets $\vec{Z'}$ of $\vec{Z}$. In words, setting any subset of variables in $\vec{W}$ to their values in $\vec{w}'$ should have no effect on φ, as long as $\vec{X}$ is kept at its current value $\vec{x}$, even if all the variables in an arbitrary subset of $\vec{Z}$ are set to their original values in the context $\vec{u}$.

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

[6] Note that we are using the word "event" here in the standard sense of "set of possible worlds" (as opposed to "transition between states of affairs"); essentially we are identifying events with propositions.

[7] We use both past tense and present tense in our examples ("was the cause" versus "is the cause"), with the usage depending on whether the scenario implied by the context $\vec{u}$ is perceived to have taken place in the past or to persist through the present.

[8] Having said that, see the end of Example 3.2 for further discussion of this issue. Disjunctive *explanations* seem more interesting, although we cannot handle them well in our framework; these are discussed in Part II.

Although we have labeled this definition "preliminary", it is actually very close to the final definition. We discuss the final definition in Section 5, after we have considered a few examples.

The core of this definition lies in AC2. Informally, the variables in $\vec{Z}$ should be thought of as describing the "active causal process" from $\vec{X}$ to φ (also called "intrinsic process" by Lewis [1986, Appendix D]).[9] These are the variables that mediate between $\vec{X}$ and φ. Indeed, we can define an *active causal process* from $\vec{X} = \vec{x}$ to φ as a minimal set $\vec{Z}$ that satisfies AC2. We would expect that the variables in an active causal process are all on a path from a variable in $\vec{X}$ to a variable in φ. This is indeed the case. Moreover, it can easily be shown that the variables in an active causal process all change their values when $(\vec{X}, \vec{W})$ is set to $(\vec{x}', \vec{w}')$ as in AC2. Any variable that does not change in this transition can be moved to W, while retaining its value in w'—the remaining variables in $\vec{Z}$ will still satisfy AC2. (See the appendix for a formal proof.) AC2(a) says that there exists a setting $\vec{x}'$ of $\vec{X}$ that changes φ to $\neg\varphi$, as long as the variables not involved in the causal process ($\vec{W}$) take on value $\vec{w}'$. AC2(a) is reminiscent of the traditional counterfactual criterion of Lewis [1973], according to which φ would be false if it were not for $\vec{X}$ being $\vec{x}$. However, AC2(a) is more permissive than the traditional criterion; it allows the dependence of φ on $\vec{X}$ to be tested under special circumstances in which the variables $\vec{W}$ are held constant at some setting $\vec{w}'$. This modification of the traditional criterion was proposed by Pearl [1998, 2000] and was named *structural contingency*—an alteration of the model M that involves the breakdown of some mechanisms (possibly emerging from external action) but no change in the context $\vec{u}$. The need to invoke such contingencies will be made clear in Example 3.2, and is further supported by the examples of Hitchcock [2001].

AC2(b), which has no obvious analogue in the literature, is an attempt to counteract the "permissiveness" of AC2(a) with regard to structural contingencies. Essentially, it ensures that $\vec{X}$ alone suffices to bring about the change from φ to $\neg\varphi$; setting $\vec{W}$ to $\vec{w}'$ merely eliminates spurious side effects that tend to mask the action of $\vec{X}$. It captures the fact that setting $\vec{W}$ to $\vec{w}'$ does not affect the causal process by requiring that changing the values of any subset of the variables in $\vec{W}$ from $\vec{w}$ to $\vec{w}'$ has no effect on the value of φ.[10] Moreover, although the values in the variables $\vec{Z}$ involved in the causal process may be perturbed by the change, the perturbation has no impact on the value of φ. The upshot of this requirement is that we are not at liberty to conduct the counterfactual test of AC2(a) under an arbitrary alteration of the model. The alteration considered must not affect the causal process. Clearly, if the contingencies considered are limited to "freezing" variables at their actual value (a restriction used by Hitchcock [2001]), so that $(M, \vec{u}) \models \vec{W} = \vec{w}'$, then AC2(b) is satisfied automatically. However, as the examples below show, genuine causation may sometimes be revealed only through a broader class of counterfactual tests in which variables in $\vec{W}$ are set to values that differ from their actual values.

[9] Recently, Lewis [2000] has abandoned attempts to define "intrinsic process" formally. Pearl's "causal beam" [Pearl 2000, p. 318] is a special kind of active causal process, in which AC2(b) is expected to hold (with $\vec{Z} = \vec{z}^*$) for all settings $\vec{w}''$ of $\vec{W}$, not necessarily the setting $\vec{w}'$ used in AC2(a).

[10] This version of AC2(b) differs slightly from that in an earlier version of this paper [Halpern and Pearl 2001]. See Appendix A.2 for more discussion of this issue.

Pearl [2000] defines a notion of *contributory cause* in addition to actual cause. Roughly speaking, if AC2(a) holds only with $\vec{W} = \vec{w}' \neq \vec{w}$, then $\vec{X} = \vec{x}$ is a contributory cause of φ; actual causality holds only if $\vec{W} = \vec{w}$. Interestingly, in all our examples in Section 4, changing $\vec{W}$ from $\vec{w}$ to $\vec{w}'$ has no impact on the value of the variables in $\vec{Z}$. That is, $(M, \vec{u}) \models [\vec{W} \leftarrow \vec{w}'](Z = z^*)$ for all $Z \in \vec{Z}$. Thus, to check AC2(b) in these examples, it suffices to show that $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}']\varphi$. We provide an example in the appendix to show that there are cases where the variables in $\vec{Z}$ can change value, so the full strength of AC2(b) is necessary.

We remark that, like the definition here, the causal beam definition [Pearl 2000] tests for the existence of counterfactual dependency in an auxiliary model of the world, modified by a select set of structural contingencies. However, whereas the contingencies selected by the beam criterion depend only on the relationship between a variable and its parents in the causal diagram, the current definition selects the modifying contingencies based on the specific cause and effect pair being tested. This refinement permits our definition to avoid certain pitfalls (see Example A.3) that are associated with graphical criteria for actual causation. In addition, the causal beam definition essentially adds another clause to AC2, placing even more stringent requirements on causality. Specifically, it requires

AC2(c). $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}'']\varphi$ for all settings $\vec{w}''$ of $\vec{W}$.

AC2(c) says that setting $\vec{X}$ to $\vec{x}$ is enough to force φ to hold, independent of the setting of $\vec{W}$.[11] We say that $\vec{X} = \vec{x}$ *strongly causes* φ if AC2(c) holds in addition to all the other conditions. As we shall see, in many of our examples, causality and strong causality coincide. In the cases where they do not coincide, our intuitions suggest that strong causality is too strong a notion.

AC3 is a minimality condition. Heckerman and Shachter [1995] have a similar minimality requirement; Lewis [2000] mentions the need for minimality as well. Interestingly, in all the examples we have considered, AC3 forces the cause to be a single conjunct of the form $X = x$. Although it is far from obvious, Eiter and Lukasiewicz [2002] and, independently, Hopkins [2001], have shown that this is in fact a consequence of our definition. However, it depends crucially on our assumption that the set $\mathcal{V}$ of endogenous variables is finite; see the appendix for further discussion of this issue. As we shall see, it also depends on the fact that we are using causality rather than strong causality.

How reasonable are these requirements? One issue that some might find inappropriate is that we allow $X = x$ to be a cause of itself. While we do not find such trivial causality terribly bothersome, it can be avoided by requiring that $\vec{X} = \vec{x} \wedge \neg\varphi$ be consistent for $\vec{X} = \vec{x}$ to be a cause of φ. More significantly, is it appropriate to invoke structural changes in the definition of actual causation? The following example may help illustrate why we believe it is.

Example 3.2: Suppose that two arsonists drop lit matches in different parts of a dry forest, and both cause trees to start burning. Consider two scenarios. In the first, called the *disjunctive scenario*, either match by itself suffices to burn down the whole forest. That is, even if only

[11] Pearl [2000] calls this invariance *sustenance*.

one match were lit, the forest would burn down. In the second scenario, called the *conjunctive scenario*, both matches are necessary to burn down the forest; if only one match were lit, the fire would die down before the forest was consumed. We can describe the essential structure of these two scenarios using a causal model with four variables:

- an exogenous variable U that determines, among other things, the motivation and state of mind of the arsonists. For simplicity, assume that $\mathcal{R}(U) = \{u_{00}, u_{10}, u_{01}, u_{11}\}$; if $U = u_{ij}$, then the first arsonist intends to start a fire iff $i = 1$ and the second arsonist intends to start a fire iff $j = 1$. In both scenarios $U = u_{11}$.

- endogenous variables ML_1 and ML_2, each either 0 or 1, where $ML_i = 0$ if arsonist i doesn't drop the lit match and $ML_i = 1$ if he does, for $i = 1, 2$.

- an endogenous variable FB for forest burns down, with values 0 (it doesn't) and 1 (it does).

Both scenarios have the same causal network (see Figure 2); they differ only in the equation for FB. For the disjunctive scenario we have $F_{FB}(u, 1, 1) = F_{FB}(u, 0, 1) = F_{FB}(u, 1, 0) = 1$ and $F_{FB}(u, 0, 0) = 0$ (where $u \in \mathcal{R}(U)$); for the conjunctive scenario we have $F_{FB}(u, 1, 1) = 1$ and $F_{FB}(u, 0, 0) = F_{FB}(u, 1, 0) = F_{FB}(u, 0, 1) = 0$. In general, we expect that the causal model for reasoning about forest fires would involve many other variables; in particular, variables for other potential causes of forest fires such lightning and unattended campfires; here we focus on that part of the causal model that involves forest fires started by arsonists. Since for causality we assume that all the relevant facts are given, we can assume here that it is known that there were no unattended campfires and there was no lightning, which makes it safe to ignore that portion of the causal model. Denote by M_1 and M_2 the (portion of the) causal models associated with the disjunctive and conjunctive scenarios, respectively. The causal network for the relevant portion of M_1 and M_2 is described in Figure 2.

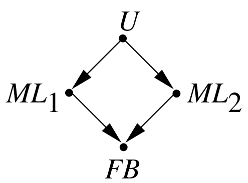

Figure 2: The causal network for M_1 and M_2.

Despite the differences in the underlying models, each of $ML_1 = 1$ and $ML_2 = 1$ is a cause of $FB = 1$ in both scenarios. We present the argument for $ML_1 = 1$ here. To show that $ML_1 = 1$ is a cause in M_1 let $\vec{Z} = \{ML_1, FB\}$, so $\vec{W} = \{ML_2\}$. It is easy to see that the contingency $ML_2 = 0$ satisfies the two conditions in AC2. AC2(a) is satisfied because, in the absence of the second arsonist ($ML_2 = 0$), the first arsonist is necessary and sufficient for the fire to occur ($FB = 1$). AC2(b) is satisfied because, if the first match is lit ($ML_1 = 1$) the contingency $ML_2 = 0$ does not prevent the fire from burning the forest. Thus, $ML_1 = 1$ is a cause of *FB* $= 1$ in M_1. (Note that we needed to set ML_2 to 0, contrary to fact, in order to reveal the latent dependence of *FB* on ML_1. Such a setting constitutes a structural change in the original model, since it involves the removal of some structural equations.)

To see that $ML_1 = 1$ is also a cause of $FB = 1$ in M_2, again let $\vec{Z} = \{ML_1, FB\}$ and $\vec{W} = \{ML_2\}$. Since $(M_2, u_{11}) \models [ML_1 \leftarrow 0, ML_2 \leftarrow 1](FB = 0)$, AC2(a) is satisfied. Moreover, since the value of ML_2 required for AC2(a) is the same as its current value (i.e., $w' = w$), AC2(b) is satisfied trivially.

This example also illustrates the need for the minimality condition $AC3$. For example, if lighting a match qualifies as the cause of fire then lighting a match and sneezing would also pass the tests of AC1 and AC2 and awkwardly qualify as the cause of fire. Minimality serves here to strip "sneezing" and other irrelevant, over-specific details from the cause.

It might be argued that allowing disjunctive causes would be useful in this case to distinguish M_1 from M_2 as far as causality goes. A purely counterfactual definition of causality would make $ML_1 = 1 \vee ML_2 = 1$ a cause of $FB = 1$ in M_1 (since, if $ML_1 = 1 \vee ML_2 = 1$ were not true, then $FB = 1$ would not be true), but would make neither $ML_1 = 1$ nor $ML_2 = 1$ individually a cause (since, for example, if $ML_1 = 1$ were not true in M_1, $FB = 1$ would still be true). Clearly, our definition does not enforce this intuition. As is well known (and as the examples in Section 4 show) purely counterfactual definitions of causality have other problems. We do not have a strong intuition as to the best way to deal with disjunction in the context of causality, and believe that disallowing it is reasonably consistent with intuitions.

This example shows that causality and strong causality do not always coincide. It is not hard to check that ML_1 and ML_2 are strong causes of *FB* in both scenarios. However, for ML_1 to be a strong cause of *FB* in the conjunctive scenario, we must include ML_2 in $\vec{Z}$ (so that $\vec{W}$ is empty); if ML_2 is in $\vec{W}$, then AC2(c) fails. Thus, with strong causality, it is no longer the case that we can take $\vec{Z}$ to consist only of variables on a path between the cause ($ML_1 = 1$ in this case) and the effect ($FB = 1$).

Moreover, suppose that we change the disjunctive scenario slightly by allowing either arsonist to have guilt feelings and call the fire department. If one arsonist calls the fire department, then the forest is saved, no matter what the other arsonist does. We can model this by allowing ML_1 and ML_2 to have a value of 2 (where $ML_i = 2$ if arsonist i calls the fire department). If either is 2, then $FB = 0$. In this situation, it is easy to check that now neither $ML_1 = 1$ nor $ML_2 = 1$ by itself is a strong cause of $FB = 1$ in the disjunctive scenario. $ML_1 = 1 \wedge ML_2 = 1$ is a cause, but it seems strange that in the disjunctive scenario, we should need to take both arsonists dropping a lit match to (strongly) cause the fire, just because we allow for the possibility that an arsonist can call the fire department. Note that this also shows that, in general, strong causes are not always single conjuncts. ∎

This is a good place to illustrate the need for structural contingencies in the analysis of actual causation. The reason we consider $ML_1 = 1$ to be a cause of $FB = 1$ in M_1 is that *if* ML_2 had been 0, rather than 1, *FB* would depend on ML_1. In words, we imagine a situation in which the second match is not lit, and we then reason counterfactually that the forest would not have burned down if it were not for the first match.

Although $ML_1 = 1$ is a cause of $FB = 1$ in both the disjunctive and conjunctive scenarios, the models M_1 and M_2 differ in regard to explanation, as we shall see in Part II of this paper. In the disjunctive scenario, the lighting of one of the matches constitutes a reasonable explanation of the forest burning down; not so in the conjunctive scenario. Intuitively, we feel that if both matches are needed for establishing a forest fire, then both $ML_1 = 1$ and $ML_2 = 1$ together would be required to fully explain the unfortunate fate of the forest; pointing to just one of these events would only beg another "How come?" question, and would not stop any serious investigating team from continuing its search for a more complete answer.

Finally, a remark concerning a *contrastive* extension to the definition of cause. When seeking a cause of φ, we are often not just interested the occurrence versus nonoccurrence of φ, but also the manner in which φ occurred, as opposed to some alternative way in which φ could have occurred [Hitchcock 1996]. We say, for example, "$X = x$ caused a fire in June as opposed to a fire in May." If we assume that there is only enough wood in the forest for one forest fire, the two contrasted events, "fire in May" and "fire in June", exclude but do not complement each other (e.g., neither rules out a fire in April.) Definition 3.1 can easily be extended to accommodate *contrastive causation*. We define "x caused φ, as opposed to φ'", where φ and φ' are incompatible but not exhaustive, by simply replacing $\neg\varphi$ with φ' in condition AC2(a) of the definition.

Contrast can also be applied to the antecedent, as in "Susan's running rather than walking to music class caused her fall." There are actually two interpretations of this statement. The first is that Susan's running is a cause of her falling; moreover, had she walked, then she would not have fallen. The second is that, while Susan's running is a cause of her falling, Susan's walking also would have caused her to fall, but she did not in fact walk. We can capture both interpretations of "$X = x$, rather than $X = x'$ for some value $x' \neq x$, caused φ (in context $\vec{u}$ in structure M)". The first is (1) $X = x$ is a cause of φ in $(M, \vec{u})$ and (2) $(M, \vec{u}) \models [X \leftarrow x']\neg\varphi$; the second is (1$'$) $X = x$ is a cause of φ in $(M, \vec{u})$ and (2$'$) AC2(b) holds for $X = x'$ and φ. That is, the only reason that $X = x'$ is not the cause of φ is that $X = x'$ is not in fact what happened in the actual world.[12] (More generally, we can make sense of "$X = x$ rather than $Y = y$ caused φ".) Contrasting both the antecedent and the consequent components is straightforward, and allows us to interpret sentences of the form: "Susan's running rather than walking to music class caused her to spend the night in the hospital, as opposed to her boyfriend's apartment."

4 Examples

In this section we show how our definition of actual causality handles some examples that have caused problems for other definitions.

Example 4.1: The first example is due to Bennett (and appears in [Sosa and Tooley 1993,

[12] As Christopher Hitchcock [private communication, 2000] has pointed out, one of the roles of such contrastive statements is to indicate that $\mathcal{R}(X)$, the set of possible values of X, should include x'. The sentence does not make sense without this assumption.

pp. 222–223]). Suppose that there was a heavy rain in April and electrical storms in the following two months; and in June the lightning took hold. If it hadn't been for the heavy rain in April, the forest would have caught fire in May. The question is whether the April rains caused the forest fire. According to a naive counterfactual analysis, they do, since if it hadn't rained, there wouldn't have been a forest fire in June. Bennett says "That is unacceptable. A good enough story of events and of causation might give us reason to accept some things that seem intuitively to be false, but no theory should persuade us that delaying a forest's burning for a month (or indeed a minute) is causing a forest fire."

In our framework, as we now show, it is indeed false to say that the April rains caused the fire, but they were a cause of there being a fire in June, as opposed to May. This seems to us intuitively right. To capture the situation, it suffices to use a simple model with three endogenous random variables:

- *AS* for "April showers", with two values—0 standing for did *not* rain heavily in April and 1 standing for rained heavily in April;
- *ES* for "electric storms", with four possible values: $(0, 0)$ (no electric storms in either May or June), (1,0) (electric storms in May but not June), (0,1) (storms in June but not May), and (1,1) (storms in both May and June);
- and F for "fire", with three possible values: 0 (no fire at all), 1 (fire in May), or 2 (fire in June).

We do not describe the context explicitly, either here or in the other examples. Assume its value $\vec{u}$ is such that it ensures that there is a shower in April, there are electric storms in both May and June, there is sufficient oxygen, there are no other potential causes of fire (such as dropped matches), no other inhibitors of fire (alert campers setting up a bucket brigade), and so on. That is, we choose $\vec{u}$ so as to allow us to focus on the issue at hand and to ensure that the right things happened (there was both fire and rain).

We will not bother writing out the details of the structural equations—they should be obvious, given the story (at least, for the context $\vec{u}$); this is also the case for all the other examples in this section. The causal network is simple: there are edges from *AS* to F and from *ES* to F. It is easy to check that each of the following holds.

- $AS = 1$ is a cause of the June fire $(F = 2)$ (taking $\vec{W} = \{ES\}$ and $\vec{Z} = \{AS, F\}$) but not of fire $(F = 2 \vee F = 1)$. That is, April showers are not a cause of the fire, but they are a cause of the June fire.
- $ES = (1, 1)$ is a cause of both $F = 2$ and $(F = 1 \vee F = 2)$. Having electric storms in both May and June caused there to be a fire.
- $AS = 1 \wedge ES = (1, 1)$ is not a cause of $F = 2$, because it violates the minimality requirement of AC3; each conjunct alone is a cause of $F = 2$. Similarly, $AS = 1 \wedge ES = (1, 1)$ is not a cause of $(F = 1 \vee F = 2)$.

$\blacksquare$

Although we did not describe the context explicitly in Example 4.1, it still played a crucial role. If we decide that the presence of oxygen is relevant then we must take this factor out of the context and introduce it as an explicit endogenous variables. Doing so can affect the causality picture (see Example 4.3). The next example already shows the importance of choosing an appropriate granularity in modeling the causal process and its structure.

Example 4.2: The following story from [Hall 2004] is an example of *preemption*, where there are two potential causes of an event, one of which preempts the other. An adequate definition of causality must deal with preemption in all of its guises.

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

Common sense suggests that Suzy's throw is the cause of the shattering, but Billy's is not. This holds in our framework too, 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).

Again, we have a simple causal network, with edges from both ST and BT to BS. In this simple causal network, BT and ST play absolutely symmetric roles, with $BS = ST \vee BT$; there is nothing to distinguish BT from ST. Not surprisingly, both Billy's throw and Suzy's throw are classified as causes of the bottle shattering in this model.

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. The model has to be refined to express this distinction. One way is to invoke a dynamic model [Pearl 2000, p. 326]; this is discussed below. A perhaps simpler way to gain expressiveness is to allow BS to be three valued, with values 0 (the bottle doesn't shatter), 1 (it shatters as a result of being hit by Suzy's rock), and 2 (it shatters as a result of being hit by Billy's rock). We leave it to the reader to check that $ST = 1$ is a cause of $BS = 1$, but $BT = 1$ is not (if Suzy hadn't thrown but Billy had, then we would have $BS = 2$). Thus, to some extent, this solves our problem. But it borders on cheating; the answer is almost programmed into the model by invoking the relation "as a result of", which requires the identification of the actual cause.

A more useful choice is to add two new random 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.

With this addition, we can go back to *BS* being two-valued. In this model, we have the causal network shown in Figure 3, with the arrow $SH \rightarrow BH$ being inhibitory; $BH = BT \wedge \neg SH$ (that is, $BH = 1$ iff $BT = 1$ and $SH = 0$). Note that, to simplify the presentation, we have omitted the exogenous variables from the causal network in Figure 3; we do so in some of the subsequent figures as well. In addition, we have given the arrows only for the particular context of interest, where Suzy throws first. In a context where Billy throws first, the arrow would go from *BH* to *SH* rather than going from *SH* to *BH*, as it does in the figure.

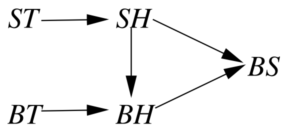

Figure 3: The rock-throwing example.

Now it is the case that $ST = 1$ is a cause of $BS = 1$. To satisfy AC2, we choose $\vec{W} = \{BT\}$ and $w' = 0$ and note that, because *BT* is *set* to 0, *BS* will track the setting of *ST*. Also note that $BT = 1$ is not a cause of $BS = 1$; there is no partition $\vec{Z} \cup \vec{W}$ that satisfies AC2. Attempting the symmetric choice $\vec{W} = \{ST\}$ and $w' = 0$ would violate AC2(b) (with $\vec{Z}' = \{BH\}$), because φ becomes false when we set $ST = 0$ and restore *BH* to its current value of 0.

This example illustrates the need for invoking subsets of $\vec{Z}$ in AC2(b). (Additional reasons are provided in Example A.3 in the appendix.) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}']\varphi$ holds if we take $\vec{Z} = \{BT, BH\}$ and $\vec{W} = \{ST, SH\}$, and thus without the requirement that AC2(b) hold for all subsets of $\vec{Z}$, $BT = 1$ would have qualified as a cause of $BS = 1$. Insisting that φ remains unchanged when both $\vec{W}$ is set to $\vec{w}'$ and $\vec{Z}'$ is set to $\vec{z}^*$ (for an arbitrary subset $\vec{Z}'$ of $\vec{Z}$) prevents us from choosing contingencies $\vec{W}$ that interfere with the active causal paths from $\vec{X}$ to φ.

This example also 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 X’s shot that killed C rather than Y’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 Y’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. This is indeed the rationale for Lewis's [1973] criterion of causation in terms of a counterfactual-dependence chain. We shall see, however, that our definition goes beyond this criterion.

It may be argued, of course, that by introducing the intermediate variables *SH* and *BH* in Hall's story we have also programmed the desired answer into the problem; after all, it is the shattering of the bottle, not *SH*, which prevents *BH*. Pearl [2000, Section 10.3.5] analyzes a similar late-preemption problem in a dynamic structural equation model, where variables are time indexed, and shows that the selection of the first action as an actual cause of the effect follows from conditions (similar to) AC1–AC3 even without specifying the owner of the hitting ball. We now present a simplified adaptation of this analysis.

Let t_1, t_2, and t_3 stand, respectively, for the time that Suzy threw her rock, the time that Billy threw his rock, and the time that the bottle was found shattered. Let H_i and BS_i be variables indicating whether the bottle is hit (H_i) and was shattered (BS_i) at time t_i (where $i = 1, 2, 3$ and $t_1 < t_2 < t_3$), with values 1 if hit (respectively, shattered), 0 if not. Roughly speaking, if we let T_i be a variable representing "someone throws the ball at time t_i" and take BS_0 to be vacuously false (i.e., always 0), then we would expect the following time-invariant equations to hold for all times t_i (not just t_1, t_2, and t_3):

$$\begin{aligned} H_i &= T_i \land \neg BS_{i-1} \\ BS_i &= BS_{i-1} \lor H_i. \end{aligned}$$

That is, the bottle is hit at time t_i if someone throws the ball at time t_i and the bottle wasn't already shattered at time t_i. Similarly, the bottle is shattered at time t_i either if it was already shattered at time t_{i-1} or it was hit at time t_i.

Since in this case we consider only times t_1, t_2, and t_3, we get the following structural equations, where we have left in the variable T_3 to bring out the essential invariance:

$$\begin{aligned} H_1 &= ST \\ BS_1 &= H_1 \\ H_2 &= BT \land \neg BS_1 \\ BS_2 &= BS_1 \lor H_2 \\ H_3 &= T_3 \land \neg BS_2 \\ BS_3 &= BS_2 \lor H_3. \end{aligned}$$

The diagram associated with this model is shown in Figure 4. In addition to these generic equations, the story also specifies that the context is such that

$$ST = 1, BT = 1, T_3 = 0.$$

The causal network in Figure 4 describes the situation.

It is not hard to show that $ST = 1$ is a cause of $BS_3 = 1$ (taking $\vec{W} = \{BT\}$ in AC2 and $w' = 0$). $BT = 1$ is not a cause of $BS_3 = 1$; it fails AC2(b) for every partition $\vec{Z} \cup \vec{W}$. To see this, note that to establish counterfactual dependence between BS_3 and BT, we must assign H_2

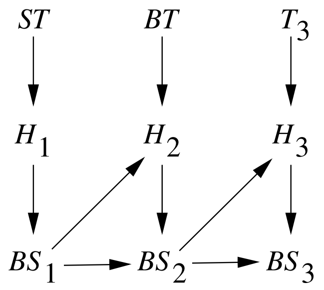

Figure 4: Time-invariant rock throwing.

to $\vec{Z}$, assign BS_1 to $\vec{W}$, and impose the contingency $BS_1 = 0$. But this contingency violates condition AC2(b), since it results in $BS_3 = 0$ when we restore H_2 to 0 (its current value).

Two features are worth emphasizing in this example. First, Suzy's throw is declared a cause of the outcome event $BS_3 = 1$ even though her throw did not hasten, delay, or change any property of that outcome. This can be made clearer by considering another outcome event, J_4 = "Joe was unable to drink his favorite chocolate cocktail from that bottle on Tuesday night." Being a consequence of BS_3, J_4 will also be classified as having been caused by Suzy's throw, not by Billy's, although J_4 would have occurred at precisely the same time and in the same manner had Suzy not thrown the ball. This implies that hastening or delaying the outcome cannot be taken as the basic principle for deciding actual causation, a principle advocated by Paul [1998].

Second, Suzy's throw is declared a cause of $BS_3 = 1$ even though there is no counterfactual dependence chain between the two (i.e., a chain $A_1 \to A_2 \to \ldots \to A_k$ where each event is counterfactually dependent on its predecessor). The existence of such a chain was proposed by Lewis [1973] as a necessary criterion for causation in cases involving preemption.[13] In the actual context, BS_2 does not depend (counterfactually) on either BS_1 or on H_2; the bottle would be shattered at time t_2 even if it were unshattered at time t_1 (since Billy's rock would have hit it), as well as if it were hit (miraculously) at time t_2. ∎

Example 4.3: Can *not* performing an action be (part of) a cause? Consider the following story, again taken from (an early version of) [Hall 2004]:

> Billy, having stayed out in the cold too long throwing rocks, contracts a serious but nonfatal disease. He is hospitalized and treated on Monday, so is fine Tuesday morning.

But now suppose the doctor does not treat Billy on Monday. Is the doctor's omission to treat Billy a cause of Billy's being sick on Tuesday? It seems that it should be, and indeed it is according to our analysis. Suppose that $\vec{u}$ is the context where, among other things, Billy is

[13] Lewis [1986, Appendix D] later amended this criterion to deal with problematic cases similar to that presented her.

sick on Monday and the situation is such that the doctor forgets to administer the medication Monday. (There is much more to the context $\vec{u}$, as we shall shortly see.) It seems reasonable that the model should have two random variables:

- *MT* for "Monday treatment", with values 0 (the doctor does not treat Billy on Monday) and 1 (he does); and
- *BMC* for "Billy's medical condition", with values 0 (recovered) and 1 (still sick).

Sure enough, in the obvious causal model, $MT = 0$ is a cause of $BMC = 1$.

This may seem somewhat disconcerting at first. Suppose there are 100 doctors in the hospital. Although only one of them was 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 fact that they didn't give him the medication also part of the cause of him still being sick on Tuesday?

In the particular model that we have constructed, the other doctors' failure to give Billy his medication is not a cause, since we have no random variables to model the other doctors' actions, just as we had no random variable in Example 4.1 to model the presence of oxygen. Their lack of action is part of the context. We factor it out because (quite reasonably) we want to focus on the actions of Billy's doctor. If we had included endogenous random variables corresponding to the other doctors, then they too would be causes of Billy's being sick on Tuesday.

With this background, we continue with Hall's modification of the original story.

> Suppose that Monday's doctor is reliable, and administers the medicine first thing in the morning, so that Billy is fully recovered by Tuesday afternoon. Tuesday's doctor is also reliable, and would have treated Billy if Monday's doctor had failed to . . . And let us add a twist: one dose of medication is harmless, but two doses are lethal.

Is the fact that Tuesday's doctor did *not* treat Billy the cause of him being alive (and recovered) on Wednesday morning?

The causal model for this story is straightforward. There are three random variables: *MT* for Monday's treatment (1 if Billy was treated Monday; 0 otherwise), *TT* for Tuesday's treatment (1 if Billy was treated Tuesday; 0 otherwise), and *BMC* for Billy's medical condition (0 if Billy is fine both Tuesday morning and Wednesday morning; 1 if Billy is sick Tuesday morning, fine Wednesday morning; 2 if Billy is sick both Tuesday and Wednesday morning; 3 if Billy is fine Tuesday morning and dead Wednesday morning). We can then describe Billy's condition as a function of the four possible combinations of treatment/nontreatment on Monday and Tuesday.

In the causal network corresponding to this causal model, shown in Figure 5, there is an edge from *MT* to *TT*, since whether the Tuesday treatment occurs depends on whether the Monday treatment occurs, and edges from both *MT* and *TT* to *BMC*, since Billy's medical condition depends on both treatments.

$$\begin{tikzcd}[column sep=small] \mathit{MT} \arrow[rr] \arrow[dr] & & \mathit{TT} \arrow[dl] \\ & \mathit{BMC} & \end{tikzcd}$$

Figure 5: Billy's medical condition.

In this causal model, it is true that $MT = 1$ is a cause of $BMC = 0$, as we would expect—because Billy is treated Monday, he is not treated on Tuesday morning, and thus recovers Wednesday morning. $MT = 1$ is also a cause of $TT = 0$, as we would expect, and $TT = 0$ is a cause of Billy's being alive ($BMC = 0 \vee BMC = 1 \vee BMC = 2$). However, $MT = 1$ is *not* a cause of Billy's being alive. It fails condition AC2(a): setting $MT = 0$ still leads to Billy's being alive (with $W = \emptyset$). Note that it would not help to take $\vec{W} = \{TT\}$. For if $TT = 0$, then Billy is alive no matter what *MT* is, while if $TT = 1$, then Billy is dead when *MT* has its original value, so AC2(b) is violated (with $\vec{Z}' = \emptyset$).

This shows that causality is not transitive, according to our definitions. Although $MT = 1$ is a cause of $TT = 0$ and $TT = 0$ is a cause of $BMC = 0 \vee BMC = 1 \vee BMC = 2$, $MT = 1$ is not a cause of $BMC = 0 \vee BMC = 1 \vee BMC = 2$. Nor is causality closed under *right weakening*: $MT = 1$ is a cause of $BMC = 0$, which logically implies $BMC = 0 \vee BMC = 1 \vee BMC = 2$, which is not caused by $MT = 1$.[14]

Hall [2000, 2004] discusses the issue of transitivity of causality, and suggests that there is a tension between the desideratum that causality be transitive and the desideratum that we allow causality due to the failure of some event to occur. He goes on to suggest that there are actually two concepts of causation: one corresponding to counterfactual dependence and the other corresponding to "production", whereby A causes B if A helped to produce B. Causation by production is transitive; causation by dependence is not.

Our definition certainly has some features of both counterfactual dependence and of production—AC2(a) captures some of the intuition of counterfactual dependence (if A hadn't happened then B wouldn't have happened if $\vec{W} = \vec{w}'$) and AC2(b) captures some of the features of production (A forced B to happen, even if $\vec{W} = \vec{w}'$). Nevertheless, we do not require two separate notions to deal with these concerns.

Moreover, whereas Hall attributes the failure of transitivity to a distinction between presence and absence of events, according to our definition, the requirement of transitivity causes problems whether or not we allow causality due to the failure of some event to occur. It is easy enough to construct a story whose causal model has precisely the same formal structure as that above, except that $TT = 0$ now means that the treatment was given and $TT = 1$ means it wasn't. (Billy starts a course of treatment on Monday which, if discontinued once started, is fatal ...)

[14] Lewis [2000] implicitly assumes right weakening in his defense of transitivity. For example, he says "... it is because of Black's move that Red's victory is caused one way rather than another. That means, I submit, that in each of these cases, Black's move did indeed cause Red's victory. Transitivity succeeds." But there is a critical (and, to us, unjustifiable) leap in this reasoning. As we already saw in Example 4.1, the fact that April rains cause a fire in June does *not* mean that they cause the fire.

Again, we don't get transitivity, but now it is because an event occurred (the treatment was given), not because it failed to occur.

Lewis [1986, 2000] insists that causality is transitive, partly to be able to deal with preemption [Lewis 1986]. As Hitchcock [2001] points out, our account handles the standard examples of preemption without needing to invoke transitivity, which, as Lewis's own examples show, leads to counterintuitive conclusions. ∎

Example 4.4: This example considers the problem of what Hall calls *double prevention*. Again, the story is taken from Hall [2004]:

> Suzy and Billy have grown up, just in time to get involved in World War III. Suzy is piloting a bomber on a mission to blow up an enemy target, and Billy is piloting a fighter as her lone escort. Along comes an enemy fighter plane, piloted by Enemy. Sharp-eyed Billy spots Enemy, zooms in, pulls the trigger, and Enemy's plane goes down in flames. Suzy's mission is undisturbed, and the bombing takes place as planned.

Does Billy deserve part of the cause for the success of the mission? After all, if he hadn't pulled the trigger, Enemy would have eluded him and shot down Suzy. Intuitively, it seems that the answer is yes, and the obvious causal model gives us this. Suppose we have the following random variables:

- *BPT* for "Billy pulls trigger", with values 0 (he doesn't) and 1 (he does);
- *LE* for "Enemy eludes Billy", with values 0 (he doesn't) and 1 (he does);
- *LSS* for "Enemy shoots Suzy", with values 0 (he doesn't) and 1 (he does);
- SST for "Suzy shoots target", with values 0 (she doesn't) and 1 (she does);
- *TD* for "target destroyed", with values 0 (it isn't) and 1 (it is).

The causal network corresponding to this model is just

$$BPT \rightarrow LE \rightarrow LSS \rightarrow SST \rightarrow TD.$$

In this model, $BPT = 1$ is a cause of $TD = 1$. Of course, $SST = 1$ is a cause of $TD = 1$ as well. It may be somewhat disconcerting to observe that $BPT = 1$ is also a cause of $SST = 1$. It seems strange to think of Billy being a cause of Suzy doing something she was planning to do all along. Part of the problem is that, according to our definition (and all other definitions of causality that we are aware of), if A enables B, then A is a cause of B. Arguably another part of the problem with $BPT = 1$ being a cause of $SST = 1$ and $TD = 1$ is that it seems to leave Suzy out of the picture altogether. We can bring Suzy more into the picture by having a random variable corresponding to Suzy's plan or intention. Suppose that we add a random

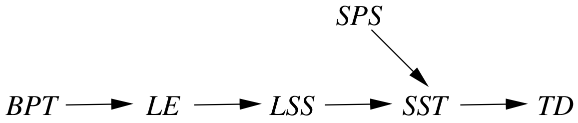

Figure 6: Blowing up the target.

variable *SPS* for “Suzy plans to shoot the target”, with values 0 (she doesn’t) and 1 (she does). Assuming that Suzy shoots if she plans to, we then get the following causal network, where now *SST* depends on both *LSS* and *SPS*:

In this case, it is easy to check that each of $BPT = 1$ and $SPS = 1$ is a cause of $TD = 1$.

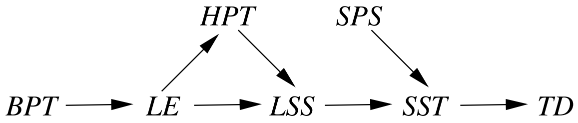

Figure 7: Blowing up the target (refined version).

Hall suggests that further complications arise if we add a second fighter plane escorting Suzy, piloted by Hillary. Billy still shoots down Enemy, but if he hadn’t, Hillary would have. The natural way of dealing with this is to add just one more variable *HPT* representing Hillary’s pulling the trigger iff $LE = 1$ (see Figure 7), but then, using the naive counterfactual criterion, one might conclude that the target will be destroyed ($TD = 1$) regardless of Billy’s action, and $BPT = 1$ would lose its “actual cause” status (of $TD = 1$). Fortunately, our definition goes beyond this naive criterion and classifies $BPT = 1$ as a cause of $TD = 1$, as expected. This can be seen by noting that the partition $\vec{Z} = \{BPT, LE, LSS, SST, TD\}$, $\vec{W} = \{HPT, SPS\}$ satisfies conditions AC1–AC3 (with w' such that $HPT = 0$ and $SPS = 1$). The intuition rests, again, on structural contingencies; although Billy’s action seems superfluous under ideal conditions, it becomes essential under a contingency in which Hillary would fail in her mission to shoot Enemy. This contingency is represented by setting *HPT* to 0 (in testing AC2(a)), irrespective of *LE*. ∎

5 A More Refined Definition

We labeled our definition “preliminary”, suggesting that there are some situations it cannot deal with. The following example illustrates the problem.

Example 5.1: Consider Example 4.2, where both Suzy and Billy throw a rock at a bottle, but Suzy’s hits first. Now suppose that there is a noise which causes Suzy to delay her throw slightly, but that she still throws before Billy. Suppose that we model this situation using the approach described in Figure 4, adding three extra variables, N (where $N = 0$ if there is no noise and $N = 1$ if there is a noise), $H_{1.5}$ (which is 1 if the bottle is hit at time $t_{1.5}$, where

$t_1 < t_{1.5} < t_2$, and 0 otherwise) and $BS_{1.5}$ (which is 1 if the bottle is shattered at time $t_{1.5}$ and 0 otherwise). In the actual situation, there is a noise and the bottle shatters at $t_{1.5}$, so $N = 1$, $H_{1.5} = 1$, and $BS_{1.5} = 1$. Just as in Example 4.2, we can show that Suzy's throw is a cause of the bottle shattering and Billy's throw is not. Not surprisingly, $N = 1$ is a cause of $BS_{1.5} = 1$ (without the noise, the bottle would have shattered at time 1). Somewhat disconcertingly though, $N = 1$ is also a cause of the bottle shattering. That is, $N = 1$ is a cause of $BS_3 = 1$.

This seems unreasonable. Intuitively, the bottle would have shattered whether or not there had been a noise. However, this intuition is actually not correct in our causal model. Consider the contingency where Suzy's throw hits the bottle. If $N = 1$ and $BS_1 = 0$, then the bottle does shatter at $t_{1.5}$. Given this, it easily follows that, according to our definition, $N = 1$ is a cause of $BS_3 = 1$.[15] ∎

The problem here is caused by what might be considered an extremely unreasonable scenario: If $N = 1$ and $BS_1 = 0$, the bottle does not shatter despite being hit by Suzy's rock. Do we want to consider such scenarios? That is up to the modeler. Intuitively, if we allow such scenarios, then the noise ought to be a cause; if not, then it shouldn't.

It is easy to modify our preliminary definition so as to be able to capture this intuition. We take an *extended causal model* to now be a tuple $(\mathcal{S}, \mathcal{F}, \mathcal{E})$, where $(\mathcal{S}, \mathcal{F})$ is a causal model, and $\mathcal{E}$ is a set of *allowable settings* for the endogenous variables. That is, if the endogenous variables are $X_1, \ldots, X_n$, then $(x_1, \ldots, x_n) \in \mathcal{E}$ if $X_1 = x_1, \ldots, X_n = x_n$ is an allowable setting. We say that a setting of a subset of the endogenous variables is allowable if it can be extended to a setting in $\mathcal{E}$. We then slightly modify clauses AC2(a) and (b) in the definition of causality to restrict to allowable settings. In the special case where $\mathcal{E}$ consists of all settings, this definition reduces to the definition we gave in Section 3. We can deal with Example 5.1 in extended causal models by disallowing settings where $BS_1 = 0 \wedge H_1 = 1$. This essentially puts us back in the original setting. The following example further illustrates the need to be able to deal with "unreasonable" settings.

Example 5.2: Fred has his finger severed by a machine at the factory ($FS = 1$). Fortunately, Fred is covered by a health plan. He is rushed to the hospital, where his finger is sewn back on. A month later, the finger is fully functional ($FF = 1$). In this story, we would not want to say that $FS = 1$ is a cause of $FF = 1$ and, indeed, according to our definition, it is not, since $FF = 1$ whether or not $FS = 1$ (in all contingencies satisfying AC2(b)).

However, suppose we introduce a new element to the story, representing a nonactual structural contingency: Larry the Loanshark may be waiting outside the factory with the intention of cutting off Fred's finger, as a warning to him to repay his loan quickly. Let *LL* represent whether or not Larry is waiting and let *LC* represent whether Larry cuts off Fred's finger. If Larry cuts off Fred's finger, he will throw it away, so Fred will not be able to get it sewn back on. In the actual situation, $LL = LC = 0$; Larry is not waiting and Larry does not cut off

[15] We thank Chris Hitchcock for bringing this example to our attention.

Fred's finger. So, intuitively, there seems to be no harm in adding this fanciful element to the story. Or is there? Suppose that, if Fred's finger is cut off in the factory, then Larry will not be able to cut off the finger himself (since Fred will be rushed off to the hospital). Now $FS = 1$ becomes a cause of $FF = 1$. For in the structural contingency where $LL = 1$, if $FS = 0$ then $FF = 0$ (Larry will cut off Fred's finger and throw it away, so it will not become functional again). Moreover, if $FS = 1$, then $LC = 0$ and $FF = 1$, just as in the actual situation.[16]

If we really want to view Larry's cutting off Fred's finger as totally fanciful, then we simply disallow all settings where $LL = 1$. On the other hand, if having fingers cut off in a way that they cannot be put on again is rather commonplace, then it seems more reasonable to view the accident as a cause of Fred's finger being functional a month after the accident. ∎

In extended models, it is also straightforward to deal with problems of causation by omission.

Example 5.3: Hall and Paul [2003] give an example due to Sarah McGrath suggesting that there may be a difference between causation by omission and causation by commission:

> Suppose Suzy goes away on vacation, leaving her favorite plant in the hands of Billy, who has promised to water it. Billy fails to do so. The plant dies—but would not have, had Billy watered it. ...Billy's failure to water the plant caused its death. But Vladimir Putin also failed to water Suzy's plant. And, had he done so, it would not have died. Why do we also not count his omission as a cause of the plant's death?

Billy is clearly a cause in the obvious structural model. So is Vladimir Putin, if we do not disallow any settings and include Putin watering the plant as one of the endogenous variables. However, if we simply disallow the setting where Vladimir Putin waters the plant. then Billy's failure to water the plants is a cause, and Putin's failure is not. We could equally well get this result by not taking Putin's watering the plant as one of the endogenous variables in the model. (Indeed, we suspect that most people modeling the problem would not include this as a random variable.)

Are we giving ourselves too much flexibility here? We believe not. It is up to a modeler to defend her choice of model. A model which does not allow us to consider Putin watering the plant can be defended in the obvious way: that is a scenario too ridiculous to consider. On the other hand, if Suzy's sister Maggie (who has a key to the house) also came by to check up on things, then it does not seem so unreasonable for Suzy to get slightly annoyed at Maggie for not watering the plant, even if she was not supposed to be the one responsible for it. Intuitively, it seems reasonable not to disallow the setting where Maggie waters the plant. ∎

[16]We thank Eric Hiddleston for bringing this example to our attention. The example is actually a variant of one originally due to Kvart [1991], although Kvart's example did not include Larry the Loanshark and was intended to show a violation of transitivity.

Considering only allowable settings plays a more significant role in our framework than just that of allowing us to ignore fanciful scenarios. As the following example shows, it helps clarify the relationship between various models of a story.

Example 5.4: This example concerns what Hall calls the distinction between causation and determination. Again, we quote Hall [2000]:

> The engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the right-hand track, instead of the left. Since the tracks reconverge up ahead, the train arrives at its destination all the same ...

Again, our causal model gets this right. Suppose we have three random variables:

- F for "flip", with values 0 (the engineer doesn't flip the switch) and 1 (she does);
- T for "track", with values 0 (the train goes on the left-hand track) and 1 (it goes on the right-hand track); and
- A for "arrival", with values 0 (the train does not arrive at the point of reconvergence) and 1 (it does).

Now it is easy to see that flipping the switch ($F = 1$) causes the train to go down the left-hand track ($T = 0$), but does not cause it to arrive ($A = 1$), thanks to AC2(a)—whether or not the switch is flipped, the train arrives.

However, our proposal goes one step beyond this simple picture. Suppose that we model the tracks using *two* variables:

- *LT* for "left-track", with values 1 (the train goes on the left-hand track) and 0 (it does not go on the left-hand track); and
- *RT* for "right-track", with values 1 (the train goes on the right-hand track) and 0 (it does not go on the right-hand track).

The resulting causal diagram is shown in Figure 8; it is isomorphic to a class of problems that Pearl [2000] calls "switching causation". It seems reasonable to disallow settings where $RT = LT = 1$; a train cannot go down more than one track. If we do not disallow any other settings, then, lo and behold, this representation classifies $F = 1$ as a cause of A. At first sight, this may seem counterintuitive: Can a change in representation turn a non-cause into a cause?

It can and it should! The change to a two-variable model is not merely syntactic, but represents a profound change in the story. The two-variable model depicts the tracks as two independent mechanisms, thus allowing one track to be set (by action or mishap) to false (or true) without affecting the other. Specifically, this permits the disastrous mishap of flipping the

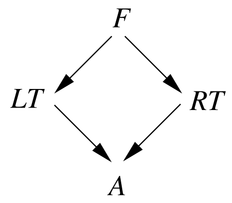

Figure 8: Flipping the switch.

switch while the left track is malfunctioning. More formally, it allows a setting where $F = 1$ and $RT = 0$. Such abnormal settings are imaginable and expressible in the two-variable model, but not in the one-variable model. Of course, if we disallow settings where $F = 1$ and $RT = 0$, or where $F = 0$ and $LT = 0$, then we are essentially back at the earlier model. The potential for such settings is precisely what renders $F = 1$ a cause of A in the model of Figure 8.[17]

Is flipping the switch a legitimate cause of the train's arrival? Not in ideal situations, where all mechanisms work as specified. But this is not what causality (and causal modeling) are all about. Causal models earn their value in abnormal circumstances, created by structural contingencies, such as the possibility of a malfunctioning track. It is this possibility that should enter our mind whenever we decide to designate each track as a separate mechanism (i.e., equation) in the model and, keeping this contingency in mind, it should not be too odd to name the switch position a cause of the train arrival (or non-arrival). ∎

Example 5.4 gives some insight into the process of model construction. While there is no way of proving that a given model is the "right" model, it is clearly important for a model to have enough random variables to express what the modeler considers to be all reasonable situations. On the other hand, by allowing for the possibility of restricting the set of possible settings in the definition of causality, we do not penalize the modeler for inadvertently having too many possible settings.

Example 5.5: The next pair of examples were introduced by Schaffer [2000] under the name *trumping preemption*. To quote Schaffer:

> Imagine that it is a law of magic that the first spell cast on a given day match the enchantment that midnight. Suppose that at noon Merlin casts a spell (the first that day) to turn the prince into a frog, that at 6:00 PM Morgana casts a spell (the only other that day) to turn the prince into a frog, and that at midnight the prince becomes a frog.

Clearly Merlin is a cause of the enchantment. What about Morgana? There is an intuition that Merlin should be the only cause, since his spell "trumps" Morgana's. Can this be captured in a causal model?

A coarse-grained model for this story has three variables:

---

[17]This can be seen by noting that condition AC2 is satisfied by the partition $\vec{Z} = \{F, LT, A\}, \vec{W} = \{LT\}$, and choosing w' as the setting $LT = 0$.

- *Mer*, with values 0 (Merlin did not cast a spell), 1 (Merlin cast a prince-to-frog spell in the morning), and 2 (Merlin cast a prince-to-frog spell in the evening);[18]
- *Mor*, with values 0, 1, 2, with interpretations similar to those for *Mer*;
- F, the outcome, with values 0 (prince) or 1 (frog).

In this model, with the obvious structural equations, both Merlin's spell and Morgana's spell are the causes of the transmogrification. (We do need to specify what happens if both Merlin and Morgana cast a spell at the same time. The choice does not affect the analysis.) The problem, of course, is that the model does not capture how Merlin's spell trumps Morgana's; Merlin and Morgana are being treated completely symmetrically. In particular, the model fails to represent the temporal precedence requirement that "the first spell on a given day match the enchantment that midnight".

To prevent Morgana's spell from being a cause, we can use a model similar in spirit to that used in the rock-throwing example. We need two additional variables, *MerE* (for Merlin's spell effective) and *MorE* (for Morgana's spell effective). The picture is very similar to Figure 3, with *MerE* and *MorE* replacing *SH* and *BH*:

$$\begin{tikzcd}
Mer \arrow[d] & & Mor \arrow[d] \\
MerE \arrow[rr] \arrow[dr] & & MorE \arrow[dl] \\
& F &
\end{tikzcd}$$

Figure 9: Merlin and Morgana.

(Again, we are not specifying what happens if Merlin and Morgana throw at the same time, because it is irrelevant to the analysis.) In this model Morgana's spell is not a cause; it fails AC2(b). This again emphasizes the point that causality is relative to a model. It is up to the modeler to ensure that the structural equations properly represent the dynamics in the story.

The second example of trumping preemption is actually due to Bas van Fraassen. Quoting Schaffer again:

> Imagine that ... the major and the sergeant stand before the corporal, both shout "Charge!" at the same time, and the corporal decides to charge.

Schaffer (and Lewis [2000]) claim that, because orders from higher-ranking soldiers trump those of lower-ranking soldiers, this is again a case of trumping preemption: the major is a cause of the charge; the sergeant is not.

[18]The variable could take on more values, allowing for other spells that Merlin could cast and other times he could cast them, but this would not affect the analysis.

Our intuition does not completely agree with that of Schaffer and Lewis in this example. In what seems to us the most obvious model of the story, both the sergeant's order and the major's order are the causes of the advance. Consider the model described in Figure 10. Assume for definiteness that the sergeant and the major can each either order an advance, order a retreat, or do nothing. Thus, M and S can each take three values, 1, -1, or 0, depending on what they do. A describes what the solders do; as the story suggests, $A = M$ if $M \neq 0$; otherwise $A = S$. In the actual context, $M = S = A = 1$. In this model, it is easy to see that both $M = 1$ and $S = 1$ are causes of $A = 1$, although $M = 1$ is a strong cause of $A = 1$, while $S = 1$ is not.

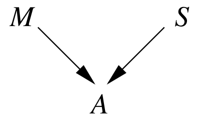

Figure 10: A simple model of the sergeant and the major

Of course, it is possible to get a model of the story that arguably captures trumping by modeling the fact that if the major actually issues an order, then the sergeant is ignored. To do this, we add a new variable *SE* that captures the sergeant's "effective" order. If the major does not issue any orders (i.e., if $M = 0$), then *SE* $= S$. If the major does issue an order, then *SE* $= 0$; the sergeant's order is effectively blocked. In this model, illustrated in Figure 11, $A = M$ if $M \neq 0$; otherwise, $A =$ *SE*.

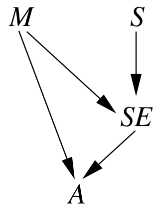

Figure 11: A model of the sergeant and the major that captures trumping.

In this model, the major does cause the corporal to advance, but the sergeant does not. For suppose we want to argue that $S = 1$ causes $A = 1$. The obvious thing to do is to take $\vec{W} = \{M\}$ and $\vec{Z} = \{S, SE, A\}$. However, this choice does not satisfy AC2(b), since if $M = 0$, *SE* $= 0$ (its original value), and $S = 1$, then $A = 0$, not 1. We leave it to the reader to check that it does not help to put *SE* into $\vec{W}$. The key point is that this more refined model allows a setting where $M = 0$, $S = 1$, and $A = 0$ (because *SE* $= 0$). That is, despite the sergeant issuing an order to attack and the major being silent, the corporal does nothing (intuitively, because of some perceived "interference" from the major, despite the major being silent).

Schaffer [2000, p. 175–176] seems to want to disallow this model, or at least to allow other mechanisms for trumping. There may well be other mechanisms for trumping. We believe that if they are spelled out carefully, it should also be possible to capture them using an appropriate causal model. However, we cannot speak about trumping preemption in our framework without being explicit as to how the trumping takes place.[19]

It is important to note that the diversity of answers in these examples does not reflect undisciplined freedom to tinker with the model so as to get the desired answer. Quite the contrary; it reflects an ambiguity in the original specification of the story, which our definition helps disambiguate. Each of the models considered reflects a legitimate interpretation of the story in terms of a distinct model of the corporal's attention-focusing strategy. For example, Figure 10 describes the corporal's strategy as a single input-output mechanism, with no intermediate steps. Figure 11 refines that model into a two-step process where the corporal first determines whether the major is silent or speaking and, in the latter case, follow the major's command. Naturally, the major should be deemed the cause of advancing (in our scenario) given this strategy. We can also imagine a completely different strategy where the sergeant, not the major, will be deemed the cause of advancing. If the corporal first determines whether or not there is conflict between the two commanders and then, in case of no conflict, pays full attention to the sergeant (perhaps because his dialect is clearer, or his posture less intimidating), it would make perfect sense then to say that the sergeant was the cause of advancing. Structural-equation models provide a language for formally representing these fine but important distinctions, and our definition translates these distinctions into different classifications of actual causes. ∎

Example 5.6: Consider an example originally due to McDermott [1995], and also considered by Collins [2000], Lewis [2000], and Hitchcock [2001]. A ball is caught by a fielder. A little further along its path there is a solid wall and, beyond that, a window. Does the fielder's catch cause the window to remain unbroken? As Lewis [2000] says,

> We are ambivalent. We can think: Yes—the fielder and the wall between them prevented the window from being broken, but the wall had nothing to do with it, since the ball never reached the wall; so it must have been the fielder. Or instead we can think: No—the wall kept the window safe regardless of what the fielder did or didn't do.

Lewis argues that our ambivalence in this case ought to be respected, and both solutions should be allowed. We can give this ambivalence formal expression in our framework. If we make both the wall and the fielder endogenous variables, then the fielder's catch is a cause of the window being safe, under the assumption that the fielder not catching the ball and the wall not being there is considered a reasonable scenario. Note that if we also have a variable for whether the ball hit the wall, then the presence of the wall is not a cause for the window's being safe in this case; the analysis is essentially the same as that of the Suzy-Billy rock-throwing example in Figure 3.[20] On the other hand, if we take it for granted the wall's presence (either by making the wall an exogenous variable, not including it in the model, or not allowing situations

[19] Of course, we could extend the framework to allow epistemic considerations, using a standard possible-worlds framework, where the "worlds" are causal models. An agent could then be uncertain about how the trumping takes place, while still knowing that the major is the cause of charge, not the sergeant. Nevertheless, in each of the possible worlds, the trumping mechanism would still have to be specified.

[20] We thank Chris Hitchcock for making this point.

where it doesn't block the ball if the fielder doesn't catch it), then the fielder's catch is not a cause of the window being safe. It would remain safe no matter what the fielder did, in any structural contingency.

This example again stresses the importance of the choice of model, and thinking through what we want to vary and what we want to keep fixed. (Much the same point is made by Hitchcock [2001].) ∎

This is perhaps a good place to compare our approach with that of Yablo [2002]. The approaches have some surface similarities. They both refine the standard notion of counterfactual dependence. We consider counterfactual dependence under some (possibly counterfactual) contingency. Yablo considers counterfactual dependence under the assumption that some feature of (or events in) the actual world remains fixed. The problem is, as Yablo himself shows, that for any $\vec{X} = \vec{x}$ and φ that actually happens, we can find some feature of the world that we can hold fixed such that φ depends on $\vec{X} = \vec{x}$. Take ψ to be the formula $\vec{X} = \vec{x} \Leftrightarrow \varphi$. If $\vec{X} = \vec{x}$ and φ are both true in the actual situation, then so is ψ. Moreover, under the assumption that ψ holds, φ depends counterfactually on $\vec{X} = \vec{x}$. In the closest world to the actual world where $\vec{X} = \vec{x} \wedge \psi$ holds, φ must hold, while in the closest world to the actual world where $\vec{X} \neq \vec{x} \wedge \psi$ holds, $\neg\varphi$ must hold. To counteract such difficulties, Yablo imposes a requirement of "naturalness" on what can be held fixed. With these requirement, a more refined notion of causation is that $\vec{X} = \vec{x}$ is a cause of φ if there is some ψ true in the actual world that can be held fixed so as to make φ counterfactually depend on $\vec{X} = \vec{x}$, and no other "more natural" ψ' can be found that makes the dependence "artificial". While Yablo does give some objective criteria for naturalness, much of the judgment is subjective, and it is not clear how to model it formally. In other words, it is not clear what relationships among variables and events must be encoded in the model in order to formally decide whether one event is "more natural" than another, or whether no other "more natural" event can be contrived. The analogous decisions in our formulation are managed by condition AC2(b), which distinguishes unambiguously between admissible and inadmissible contingencies. In addition, it restricts the form of contingencies; only contingencies of the form $\vec{W} = \vec{w}$ are allowed, and not, for example, contingencies such as $X = Y$.

6 Discussion

We have presented a formal representation of causal knowledge and a principled way of determining actual causes from such knowledge. We have shown that the counterfactual approach to causation, in the tradition of Hume and Lewis, need not be abandoned; the language of counterfactuals, once supported with structural semantics, can yield a plausible and elegant account of actual causation that resolves major difficulties in the traditional account.

The essential principles of our account include

- using structural equations to model causal mechanisms and counterfactuals;

- using uniform counterfactual notation to encode and distinguish facts, actions, outcomes, processes, and contingencies;
- using structural contingencies to uncover latent counterfactual dependencies;
- careful screening of these contingencies to avoid tampering with the causal processes to be uncovered.

Our approach also stresses the importance of careful modeling. In particular, it shows that the choice of model granularity can have a significant effect on the causality relation. This perhaps can be viewed as a deficiency in the approach. We prefer to think that it shows that the internal structures of the processes assumed to underlie causal stories play a crucial role in our judgment of actual causation, and that it is important therefore to properly cast such stories in a language that represents those structures explicitly. Our approach is built on just such a language.

As the examples have shown, much depends on choosing the "right" set of variables with which to model a situation, which ones to make exogenous, and which to make endogenous. While the examples have suggested some heuristics for making appropriate choices, we do not have a general theory for how to make these choices. We view this as an important direction for future research. (See [Hitchcock 2003] for some preliminary discussion of the issue of finding "good" models.)

While we do feel that it should be possible to delineate good guidelines for constructing appropriate models, ultimately, the choice of model is a subjective one. The choice of which variables to focus on and which to ignore (that is, the choice of exogenous and endogenous variables) and the decision as to which contingencies to take seriously (that is, which settings to take as allowable) is subjective, and depends to some extent on what the model is being used for. (This issue arises frequently in discussions of causality and the law [Hart and Honoré 1985].) By way of contrast, most of the work in the philosophy literature seems to implicitly assume that, in any given situation, there is one correct answer as to whether A is a cause of B. Rather than starting with a model, there are assumed to be events in the world; new events can be created to some extent as needed, leading to issues like "fragility" of events and how fine-grained events should be (see, for example, [Lewis 2000; Paul 2000]).

of causality is "right". However, the fact that it deals so well with the many difficult examples in the literature does provide some support for the reasonableness of the definition. Further support is provided by the ease with which it can be extended to define other notions, such as explanation (see Part II of this paper) and responsibility and blame [Chockler and Halpern 2004].

A Appendix: Some Technical Issues

In this appendix, we consider some technical issues related to the definition of causality.

A.1 The active causal process

We first show that, without loss of generality, the variables in the set $\vec{Z}$ in condition AC2 of the definition of causality can all be taken to be on a path from a variable in $\vec{X}$ to a variable in φ. In fact, they can, without loss of generality, be assumed to change value when $\vec{X}$ is set to $\vec{x}'$ and $\vec{W}$ is set to $\vec{w}'$. More formally, consider the following strengthening of AC2:

AC2′. There exists a partition $(\vec{Z}, \vec{W})$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w}')$ of the variables in $(\vec{X}, \vec{W})$ such that, if $(M, \vec{u}) \models Z = z^*$ for $Z \in \vec{Z}$, then

(a) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}'](\neg\varphi \wedge Z \neq z^*)$ for all $Z \in \vec{Z}$;

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

As we now show, we could have replaced AC2 by AC2′; it would not have affected the notion of causality. Say that $\vec{X} = \vec{x}$ is an *actual cause′* of φ if AC1, AC2′, and AC3 hold.

Proposition A.1: *$\vec{X} = \vec{x}$ is an actual cause of φ iff $\vec{X} = \vec{x}$ is an actual cause′ of φ.*

Proof: The "if" direction is immediate, since AC2′ clearly implies AC2. For the "only if" direction, suppose that $\vec{X} = \vec{x}$ is a cause of φ. Let $(\vec{Z}, \vec{W})$ be the partition of $\mathcal{V}$ and $(\vec{x}', \vec{w}')$ the setting of the variables in $(\vec{X}, \vec{W})$ guaranteed to exist by AC2. Let $\vec{Z}' \subseteq \vec{Z}$ consist of variables $Z \in \vec{Z}$ such that $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}'](Z \neq z^*)$. Let $\vec{W}' = \mathcal{V} - \vec{Z}'$. Notice that $\vec{W}'$ is a superset of $\vec{W}$. Moreover, a priori, $\vec{W}'$ may contain some variables in $\vec{X}$, although we shall show that this is not the case. Let $\vec{w}''$ be a setting of the variables in $\vec{W}$ that agrees with $\vec{w}'$ on the variables in $\vec{W}$ and for $Z \in \vec{Z} \cap \vec{W}'$, sets Z to z^* (its original value). Note that if there is a variable $V \in \vec{X} \cap \vec{W}'$, then the setting of V is the same in $\vec{x}'$, $\vec{x}$, and $\vec{w}''$. Thus, even if $\vec{X}$ and $\vec{W}'$ have a nonempty intersection, the models $M_{\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}'}$ and $M_{\vec{X} \leftarrow \vec{x}', \vec{W}' \leftarrow \vec{w}''}$ are well defined. Since $Z = z^*$ in the unique solution to the equations in $M_{\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}'}$ and the equations in $M_{\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}'}$, it follows that (a) the equations in $M_{\vec{X} \leftarrow \vec{x}', \vec{W}' \leftarrow \vec{w}''}$ and $M_{\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}'}$ have the same solutions and (b) the equations in $M_{\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}''}$ and $M_{\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}'}$ have the same solutions. Thus, $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W}' \leftarrow \vec{w}''](\neg\varphi \wedge (Z \neq z^*))$ for all $Z \in \vec{Z}'$ and $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}'](\varphi \wedge (Z = z^*))$ for all $Z \in \vec{Z}'$. That is, AC2′ (and hence AC2) holds for the pair $(\vec{Z}', \vec{W}')$. It follows that $\vec{W}' \cap \vec{X} = \emptyset$, for otherwise $\vec{X} = \vec{x}$ is not a cause of φ: it violates AC3. Thus, $\vec{Z}' \supseteq \vec{X}$, and $\vec{X} = \vec{x}$ is a cause′ of φ, as desired. ∎

Proposition A.1 shows that, without loss of generality, the variables in $\vec{Z}$ can be taken to be "active" in the causal process, in that they change value when the variables in $\vec{X}$ do. This means that each variable in $\vec{Z}$ must be a descendant of some variable in $\vec{X}$ in the causal graph. The next result shows that, without loss of generality, we can also assume that the variables in $\vec{Z}$ are on a path from a variable in $\vec{X}$ to a variable that appears in φ. Recall that we defined an active causal process to consist of a minimal set $\vec{Z}$ that satisfies AC2.

Proposition A.2: *All the variables in an active causal process corresponding to a cause $\vec{X} = \vec{x}$ for φ in $(M, \vec{u})$ must be on a path from some variable in $\vec{X}$ to a variable in φ in the causal network corresponding to M.*

Proof: Suppose that $\vec{Z}$ is an active causal process, $(\vec{Z}, \vec{W})$ is the partition satisfying AC2 using the setting $(\vec{x}', \vec{w}')$. By Proposition A.1, all the variables in $\vec{Z}$ must be descendants of a variable in $\vec{X}$. Suppose that some variable $Z \in \vec{Z}$ is not on a path from a variable in $\vec{X}$ to a variable in φ. That means there is no path from Z to a variable in φ. It follows that there is no path from Z to a variable $Z' \in \vec{Z}$ that is on a path from a variable in $\vec{X}$ to a variable in φ. Thus, changing the value of Z cannot affect the value of φ nor of any variable $Z' \in \vec{Z}$. Let $\vec{Z}' = \vec{Z} - \{Z\}$ and $\vec{W}' = \vec{W} \cup \{Z\}$. Extend $\vec{w}'$ to $\vec{w}''$ by assigning Z to its original value z^* in context $(M, \vec{u})$. It is now immediate from the preceding observations that $(\vec{Z}', \vec{W}')$ is a partition satisfying AC2 using the setting $(\vec{x}', \vec{w}'')$. This contradicts the minimality of $\vec{Z}$. ∎

A.2 A closer look at AC2(b)

Clause AC2(b) in the definition of causality is complicated by the need to check that φ remains true if $\vec{X}$ is set to $\vec{x}$, any subset of the variables in $\vec{W}$ is set to $\vec{w}'$, and all the variables in an arbitrary subset $\vec{Z}'$ of $\vec{Z}$ are set to their original values $\vec{z}^*$ (that is, the values they had in the original context, where $\vec{X} = \vec{x}$ and $\vec{W} = \vec{w}$). This check would be simplified considerably if, for each variable $z \in \vec{Z}$ and each subset $\vec{W}'$ of $\vec{W}$, we have that $Z = z^*$ when $\vec{X} = \vec{x}$ and $\vec{W}' = \vec{w}'$; that is, if we require in AC2(b) that $(M, u) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow w'](Z = z^*)$ for all variables $Z \in \vec{Z}$ and all subsets $\vec{W}'$ of $\vec{W}$. (Note that this requirement would imply the current requirement.) This stronger requirement holds in all the examples we have considered so far. However, the following example shows that it does not hold in general.

Example A.3: Imagine that a vote takes place. For simplicity, two people vote. The measure is passed if at least one of them votes in favor. In fact, both of them vote in favor, and the measure passes. This version of the story is almost identical to the disjunctive scenario in Example 3.2. If we use V_1 and V_2 to denote how the voters vote ($V_i = 0$ if voter i votes against and $V_i = 1$ if she votes in favor) and P to denote whether the measure passes ($P = 1$ if it passes, $P = 0$ if it doesn't), then in the context where $V_1 = V_2 = 1$, it is easy to see that each of $V_1 = 1$ and $V_2 = 1$ is a cause of $P = 1$. However, suppose we now assume that there is a voting machine that tabulates the votes. Let M represent the total number of votes recorded by the machine. Clearly $M = V_1 + V_2$ and $P = 1$ iff $M \geq 1$. The following causal network represents this more refined version of the story. In this more refined scenario, $V_1 = 1$ and $V_2 = 1$ are still both causes of $P = 1$. Consider $V_1 = 1$. Take $\vec{Z} = \{V_1, M, P\}$ and $\vec{W} = V_2$. Much like the simpler version of the story, if we choose the contingency $V_2 = 0$, then P is counterfactually dependent on V_1, so AC2(a) holds. To check that this contingency satisfies AC2(b), note that setting V_1 to 1 and V_2 to 0 results in $P = 1$, even if we also set M to 2 (its current value). However, if we had insisted in AC2(b) that $(M, u) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow w'](Z = z^*)$ for all variables

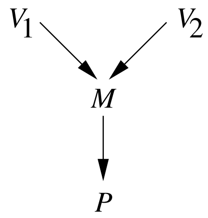

Figure 12: An example showing the need for AC2(b).

$Z \in \vec{Z}$ (which in this case means that M would have to retain its original value of 2 when $V_1 = 1$ and $V_2 = 0$), then neither $V_1 = 1$ nor $V_2 = 1$ would be a cause of $P = 1$ (although $V_1 = 1 \wedge V_2 = 1$ would be a cause of $P = 1$). Since, in general, one can always imagine that a change in one variable produces some feeble change in another, we cannot insist on the variables in $\vec{Z}$ remaining constant; instead, we require merely that changes in $\vec{Z}$ not affect φ. ∎

We remark that this example is not handled correctly by Pearl's causal beam definition. According to the causal beam definition, there is no cause for $P = 1$! It can be shown that if $X = x$ is an actual (or contributory) cause of $Y = y$ according to the causal beam definition given in [Pearl 2000], then it is an actual cause according to the definition here. As Example A.3 shows, the converse is not necessarily true.

Another complicating factor in AC2(b) is that the requirement must hold for all subsets $\vec{W}'$ of $\vec{W}$. In a preliminary version of this paper [Halpern and Pearl 2001], we required only that AC2(b) hold for $\vec{W}$. That is, the condition we had was

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

However, as Hopkins and Pearl [2002] pointed out, AC2(b′) is too permissive. To use their example, suppose that a prisoner dies either if A loads B's gun and B shoots, or if C loads and shoots his gun. Taking D to represent the prisoner's death and making the obvious assumptions about the meaning of the variables, we have that $D = 1$ iff $(A = 1 \wedge B = 1) \vee (C = 1)$. Suppose that in the actual context u, A loads B's gun, B does not shoot, but C does load and shoot his gun, so that the prisoner dies. Clearly $C = 1$ is a cause of $D = 1$. We would not want to say that $A = 1$ is a cause of $D = 1$, given that B did not shoot (i.e., given that $B = 0$). However, with AC2(b′), $A = 1$ is a cause of $D = 1$. For we can take $\vec{W} = \{B, C\}$ and consider the contingency where $B = 1$ and $C = 0$. It is easy to check that AC2(a) and AC2(b′) hold for this contingency, so under the old definition, $A = 1$ was a cause of $D = 1$. However, AC2(b) fails in this case, for $(M, u) \models [A \leftarrow 1, C \leftarrow 0]D = 0$.

A.3 Causality with infinitely many variables

Throughout this paper, we have assumed that $\mathcal{V}$, the set of exogenous variables, is finite. Our definition (in particular, the minimality clause AC3) has to be modified if we drop this assumption. To see why, consider the following example:

Example A.4: Suppose that $\mathcal{V} = \{X_0, X_1, X_2, \ldots, Y\}$. Further assume that the structural equations are such that $Y = 1$ iff infinitely many of the X_i's are 1; otherwise $Y = 0$. Suppose that in the actual context, all of the X_i's are 1 and, of course, so is Y. What is the cause of $Y = 1$?

According to our current definitions, it is actually not hard to check that there is no event which is the cause of $Y = 1$. For suppose that $\wedge_{i \in I} X_i = 1$ is a cause of $Y = 1$, for some subset I of the natural numbers. If I is finite, then to satisfy AC2(a), we must take $\vec{W}$ to be a cofinite subset of the X_i's (that is, $\vec{W}$ must include all but finitely many of the X_i's). But then if we set all but finitely many of the X_i's in $\vec{W}$ to 0 (as we must to satisfy AC2(a) if I is finite), AC2(b) fails. On the other hand, if I is infinite and there exists a partition $(\vec{Z}, \vec{W})$ such that AC2(a) and (b) hold, then if I' is the result of removing the smallest element from I, it is easy to see that $\wedge_{i \in I'} X_i = 1$ also satisfies AC2(a) and (b), so AC3 fails. ∎

Example A.4 shows that the definition of causality must be modified if $\mathcal{V}$ is infinite. It seems that the minimality condition AC3 should be modified. Here is a suggested modification:

AC3$'$. If any strict subset $\vec{X}'$ of $\vec{X}$ satisfies conditions AC1 and AC2, then there is a strict subset $\vec{X}''$ of $\vec{X}'$ that also satisfies AC1 and AC2.

It is easy to see that AC3 and AC3$'$ agree if $\mathcal{V}$ is finite. Roughly speaking, AC3$'$ says that if there is a minimal conjunction that satisfies AC1 and AC2, then it is a cause. If there is no minimal one (because there is an infinite descending sequence), then any conjunction along the sequence qualifies as a cause.

If we use AC3$'$ instead of AC3, then in Example A.4, $\wedge_{i \in I} X_i = 1$ is a cause of $Y = 1$ as long as I is infinite. Note that it is no longer the case that we can restrict to a single conjunct if $\mathcal{V}$ is infinite.

We do not have sufficient experience with this definition to be confident that it is indeed just what we want, but it seems like a reasonable choice.

A.4 Causality in nonrecursive models

We conclude by considering how the definition of causality can be modified to deal with nonrecursive models. In nonrecursive models, there may be more than one solution to an equation in a given context, or there may be none. In particular, that means that a context no longer necessarily determines the values of the endogenous variables. Earlier, we identified a primitive event such as $X = s$ with the basic causal formula $[](X = x)$, that is, with the special case of a formula of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$ with $k = 0$. $(M, \vec{u}) \models [](X = x)$ if $X = x$ in all solutions to the equations where $\vec{U} = \vec{u}$. It seems reasonable to identify $[](X = x)$ with $X = x$ if there is a unique solution to these equations. But it is not so reasonable if there may be several solutions, or no solution. What we really want to do is to be able to say that $X = x$ under a particular setting of the variables. Thus, we now take the truth of a primitive event such as $X = x$ relative not just to a context, but to a complete description $(\vec{u}, \vec{v})$ of the values of both the exogenous and the endogenous variables. That is, $(M, \vec{u}, \vec{v}) \models X = x$ if X has value x in $\vec{v}$. Since the truth of $X = x$ depends on just $\vec{v}$, not $\vec{u}$, we sometimes write $(M, \vec{v}) \models X = x$. We extend this definition to Boolean combinations of primitive events in the standard way. We then define $(M, \vec{u}, \vec{v}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M, \vec{v}') \models \varphi$ for all solutions $(\vec{u}, \vec{v}')$ to the equations in $M_{\vec{Y} \leftarrow \vec{y}}$. Since the truth of $[\vec{Y} \leftarrow \vec{y}](X = x)$ depends only on the context $\vec{u}$ and not on $\vec{v}$, we typically write $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}](X = x)$.

The formula $\langle \vec{Y} \leftarrow \vec{y} \rangle (X = x)$ is the dual of $[\vec{Y} \leftarrow \vec{y}](X = x)$; that is, it is an abbreviation of $\neg[\vec{Y} \leftarrow \vec{y}](X \neq x)$. It is easy to check that $(M, \vec{u}, \vec{v}) \models \langle \vec{Y} \leftarrow \vec{y} \rangle (X = x)$ if in some solution to the equations in $M_{\vec{Y} \leftarrow \vec{y}}$ in context $\vec{u}$, the variable X has value x. For recursive models, it is immediate that $[\vec{Y} \leftarrow \vec{y}](X = x)$ is equivalent to $\langle \vec{Y} \leftarrow \vec{y} \rangle (X = x)$, since all equations have exactly one solution.

With these definitions in hand, it is easy to state our definition of causality for arbitrary models. Note it is now taken with respect to a tuple $(M, \vec{u}, \vec{v})$, since we need the values of the exogenous variables to define the actual world.

Definition A.5: $\vec{X} = \vec{x}$ is an *actual cause of* φ *in* $(M, \vec{u}, \vec{v})$ if the following three conditions hold.

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

AC2. There exists a partition $(\vec{Z}, \vec{W})$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w}')$ of the variables in $(\vec{X}, \vec{W})$ such that if $(M, \vec{u}, \vec{v}) \models \vec{Z} = \vec{z}^*$, then

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

(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*]\varphi$ for all subsets $\vec{Z}'$ of $\vec{Z}$. (Note that in part (a) we require that the value of φ change only in some solution to the equations, while in (b), we require that it stay true in *all* solutions.)

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

While this seems like the most natural generalization of the definition of causality to deal with nonrecursive models, we have not examined examples to verify that this definition gives the expected result, partly because all the standard examples are most naturally modeled using recursive models.

Acknowledgments

We thank Christopher Hitchcock for many useful discussions, for his numerous comments on earlier versions of the paper, and for pointing out Example 5.1; Mark Hopkins, James Park, Wolfgang Spohn, and Zoltan Szabo for stimulating discussions; and Eric Hiddleston for pointing out Example 5.2.

References

Balke, A. and J. Pearl (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Proc. Tenth Conference on Uncertainty in Artificial Intelligence (UAI '94)*, pp. 46–54.

Chockler, H. and J. Y. Halpern (2004). Responsibility and blame: A structural-model approach. *Journal of A.I. Research*, 93–115.

Collins, J. (2000). Preemptive preemption. *Journal of Philosophy XCVII*(4), 223–234.

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

Eiter, T. and T. Lukasiewicz (2002). Complexity results for structure-based causality. *Artificial Intelligence 142*(1), 53–89.

Galles, D. and J. Pearl (1997). Axioms of causal relevance. *Artificial Intelligence 97*(1–2), 9–43.

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

Good, I. (1993). A tentative measure of probabilistic causation relevant to the philosophy of the law. *J. Statist. Comput. and Simulation 47*, 99–105.

Hall, N. (2000). Causation and the price of transitivity. *Journal of Philosophy XCVII*(4), 198–222.

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. and L. Paul (2003). Causation and its counterexamples: a traveler's guide. Unpublished manuscript.

Halpern, J. Y. (2000). Axiomatizing causal reasoning. *Journal of A.I. Research 12*, 317–337.

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 (2004). Causes and explanations: A structural-model approach. Part II: Explanations. *British Journal for Philosophy of Science*.

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

Heckerman, D. and R. Shachter (1995). Decision-theoretic foundations for causal reasoning. *Journal of A.I. Research 3*, 405–430.

Hitchcock, C. (1996). The role of contrast in causal and explanatory claims. *Synthese 107*, 395–419.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy XCVIII*(6), 273–299.

Hitchcock, C. (2003). Routes, processes, and chance-lowering causes. In P. Dowe and P. Noordhof (Eds.), *Cause and Chance: Causation in an Indeterministic World*. New York: Routledge.

Hopkins, M. (2001). A proof of the conjunctive cause conjecture. Unpublished manuscript.

Hopkins, M. and J. Pearl (2002). Clarifying the usage of structural models for commonsense causal reasoning. In *Proc. AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*.

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

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

Kvart, I. (1991). Transitivity and preemption of causal relevance. *Philosophical Studies LX14*, 125–160.

Lewis, D. (1973). 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. (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.

Lin, F. (1995). Embracing causality in specifying the indeterminate effects of actions. In *Proc. Fourteenth International Joint Conference on Artificial Intelligence (IJCAI '95)*, pp. 1985–1991.

McDermott, M. (1995). Redundant causation. *British Journal for the Philosophy of Science 40*, 523–544.

Michie, D. (1999). Adapting Good's Q theory to the causation of by individual events. In D. M. K. Furukawa and S. Muggleton (Eds.), *Machine Intelligence 15*, pp. 60–86. Oxford: Oxford University Press.

Paul, L. (1998). Keeping track of the time: Emending the counterfactual analysis of causation. *Analysis 3*, 191–198.

Paul, L. (2000). Aspect causation. *Journal of Philosophy XCVII*(4), 235–256.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*(4), 669–710.

Pearl, J. (1998). On the definition of actual cause. Technical Report R-259, Department of Computer Science, University of California, Los Angeles, Calif.

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

Reiter, R. (2001). *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. Cambridge, Mass.: MIT Press.

Sandewall, E. (1994). *Features and Fluents*, Volume 1. Oxford: Clarendon Press.

Schaffer, J. (2000). Trumping preemption. *Journal of Philosophy XCVII*(4), 165–181.

Sosa, E. and M. Tooley (Eds.) (1993). *Causation*. Oxford Readings in Philosophy. Oxford: Oxford University Press.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction, and Search*. New York: Springer-Verlag.

Tian, J. and J. Pearl (2000). Probabilities of causation: bounds and identification. *Annals of Mathematics and Artificial Intelligence 28*, 287–313.

Wright, R. (1988). Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts. *Iowa Law Review 73*, 1001–1077.

Yablo, S. (2002). De facto dependence. *Journal of Philosophy XCIX*(4), 130–148.