Defaults and Normality in Causal Structures

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

Abstract

A serious defect with the Halpern-Pearl (HP) definition of causality is repaired by combining a theory of causality with a theory of defaults. In addition, it is shown that (despite a claim to the contrary) a cause according to the HP condition need not be a single conjunct. A definition of causality motivated by Wright's NESS test is shown to always hold for a single conjunct. Moreover, conditions that hold for all the examples considered by HP are given that guarantee that causality according to (this version) of the NESS test is equivalent to the HP definition.

1 Introduction

Getting an adequate definition of causality is difficult. There have been numerous attempts, in fields ranging from philosophy to law to computer science (see, e.g., [Collins, Hall, and Paul 2004; Hart and Honoré 1985; Pearl 2000]). A recent definition by Halpern and Pearl (HP from now on), first introduced in [Halpern and Pearl 2001], using structural equations, has attracted some attention recently. The intuition behind this definition, which goes back to Hume [1748], is that A is a cause of B if, had A not happened, B would not have happened. For example, despite the fact that it was raining and I was drunk, the faulty brakes are the cause of my accident because, had the brakes not been faulty, I would not have had the accident. As is well known, this definition does not quite work. To take an example due to Wright [1985], suppose that Victoria, the victim, drinks a cup of tea poisoned by Paula, but before the poison takes effect, Sharon shoots Victoria, and she dies. We would like to call Sharon's shot the cause of the Victoria's death, but if Sharon hadn't shot, Victoria would have died in any case. HP deal with this by, roughly speaking, considering the contingency where Sharon does not shoot. Under that contingency, Victoria dies if Paula administers the poison, and otherwise does not. To prevent the poisoning from also being a cause of Paula's death, HP put some constraints on the contingencies that could be considered.

Unfortunately, two significant problems have been found with the original HP definition, each leading to situations where the definition does not match most people's intuitions regarding causality. The first, observed by Hopkins and Pearl [2003] (see Example 3.3), showed that the constraints on the contingencies were too liberal. This problem was dealt with in the journal version of the HP paper [Halpern and Pearl 2005] by putting a further constraint on contingencies. The second problem is arguably deeper. As examples of Hall [2007] and Hiddleston [2005] show, the HP definition gives inappropriate answers in cases that have structural equations isomorphic to ones where the HP definition gives the appropriate answer (see Example 4.1). Thus, there must be more to causality than just the structural equations. The final HP definition recognizes this problem by viewing some contingencies as "unreasonable" or "farfetched". However, in some of the examples, it is not clear why the relevant contingencies are more farfetched than others. I show that the problem is even deeper than that: there is no way of viewing contingencies as "farfetched" independent of actual contingency that can solve the problem.

This paper has two broad themes, motivated by the two problems in the HP definition. First, I propose a general approach for dealing with the second problem, motivated by the following well-known observation in the psychology literature [Kahneman and Miller 1986, p. 143]: "an event is more likely to be undone by altering exceptional than routine aspects of the causal chain that led to it." In the language of this paper, a contingency that differs from the actual situation by changing something that is atypical in the actual situation is more reasonable than one that differs by changing something that is typical in the actual situation. To capture this intuition formally, I use a well-understood approach to dealing with defaults and normality [Kraus, Lehmann, and Magidor 1990]. Combining a default theory with causality, using the intuitions of Kahnemann and Miller, leads to a straightforward solution to the second problem. The idea is that, when showing that if A hadn't happened then B would not have happened, we consider only contingencies that are more normal than the actual world. For example, if someone typically leaves work at 5:30 PM and arrives home at 6, but, due to unusually bad traffic, arrives home at 6:10, the bad traffic is typically viewed as the cause of his being late, not the fact that he left at 5:30 (rather than 5:20).

The second theme of this paper is a comparison of the HP definition to perhaps the best worked-out approach to

*Supported in part by NSF under under grants ITR-0325453 and IIS-0534064, and by AFOSR under grant FA9550-05-1-0055.

causality in the legal literature: the NESS (Necessary Element of a Sufficient Set) test, originally described by Hart and Honoré [1985], and worked out in much greater detail by Wright [1985, 1988, 2001]. This is motivated in part by the first problem. As shown by Eiter and Lukasiewicz [2002] and Hopkins [2001], the original HP definition had the property that causes were always single conjuncts; that is, it is never the case that $A \wedge A'$ is a cause of B if $A \neq A'$. This property, which plays a critical role in the complexity results of Eiter and Lukasiewicz [2002], was also claimed to hold for the revised definition [Halpern and Pearl 2005] (which was revised precisely to deal with the first problem) but, as I show here, it does not. Nevertheless, for all the examples considered in the literature, the cause is always a single conjunct. Considering the NESS test helps explain why.

While the NESS test is simple and intuitive, and deals well with many examples, as I show here, it suffers from some serious problems. In In particular, it lacks a clear definition of what it means for a set of events to be *sufficient* for another event to occur. I provide such a definition here, using ideas from the HP definition of causality. Combining these ideas with the intuition behind the NESS test leads to a definition of causality that (a) often agrees with the HP definition (indeed, does so on all the examples in the HP paper) and (b) has the property that a cause is always a single conjunct. I provide a sufficient condition (that holds in all the examples in the HP paper) for when the NESS test definition implies the HP definition, thus also providing an explanation as to why the cause is a single conjunct according to the HP definition in so many cases.

I conclude this introduction with a brief discussion on related work. There has been a great deal of work on causality in philosophy, statistics, AI, and the law. It is beyond the scope of this paper to review it; the HP paper has some comparison of the HP approach to other, particularly those in the philosophy literature. It is perhaps worth mentioning here that the focus of this work is quite different from the AI work on formal action theory (see, for example, [Lin 1995; Sandewall 1994; Reiter 2001]), which is concerned with applying causal relationships so as to guide actions, as opposed to the focus here on extracting the actual causality relation from a specific scenario.

The rest of this paper is organized as follows. In Section 2, I provide a brief introduction to structural equations and causal models, so as to make this paper self-contained. In Section 3, I review the HP definition, and show that, in general, causes are not always single conjuncts. In Section 4, I show how the HP definition can be combined with standard approaches for modeling defaults, and how that deals with the various problems that have been raised. In Section 5, I compare the structural-model definition of causality is compared to Wright's [1985, 1988, 2001] NESS test, and give a formal analogue of the NESS test combined with ideas in the HP definition. I conclude in Section 6. Proofs can be found in the appendix.

2 Causal Models

In this section, I briefly review the formal model of causality used in the HP definition. More details, intuition, and motivation can be found in [Halpern and Pearl 2005] and the references therein.

The HP approach assumes that the world is described in terms of random variables and their values. For example, if we are trying to determine whether a forest fire was caused by lightning or an arsonist, we can take the world to be described by three random variables:

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

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

Some random variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. For example, to model the fact that if a match is lit or lightning strikes then a fire starts, we could use the random variables M, FF, and L as above, with the equation $FF = \max(L, M)$. The equality sign in this equation should be thought of more like an assignment statement in programming languages; once we set the values of FF and L, then the value of FF is set to their maximum. However, despite the equality, if a forest fire starts some other way, that does not force the value of either M or L to be 1.

It is conceptually useful to split the random variables into two sets: the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. For example, in the forest fire example, the variables M, L, and FF are endogenous. However, we want to take as given that there is enough oxygen for the fire and that the wood is sufficiently dry to burn. In addition, we do not want to concern ourselves with the factors that make the arsonist drop the match or the factors that cause lightning. These factors are all determined by the exogenous variables.

Formally, a *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a *signature*, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathcal{F}$ defines a set of *modifiable structural equations*, relating the values of the variables. A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (that is, the set of values over which Y *ranges*). $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted

F_X such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V}-\{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$. This mathematical notation just makes precise the fact that F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. If there is one exogenous variable U and three endogenous variables, X, Y, and Z, then F_X defines the values of X in terms of the values of Y, Z, and U. For example, we might have $F_X(u, y, z) = u + y$, which is usually written as $X = U + Y$.[1] Thus, if $Y = 3$ and $U = 2$, then $X = 5$, regardless of how Z is set.

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

The key role of the structural equations is to define what happens in the presence of external interventions. For example, we can explain what happens if the arsonist does *not* drop the match. In the disjunctive model, there is a forest fire exactly exactly if there is lightning; in the conjunctive model, there is definitely no fire. Setting the value of some variable X to x in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model denoted $M_{X=x}$. In the new causal model, since the value of X is set, X is removed from the list of endogenous variables. That means that there is no longer an equation F_X defining X. Moreover, X is no longer an argument in the equation F_Y characterizing another endogenous variable Y. The new equation for Y is the one that results by substituting x for X. More formally, $M_{X=x} = (\mathcal{S}_X, \mathcal{F}^{X=x})$, where $\mathcal{S}_X = (\mathcal{U}, \mathcal{V} - \{X\}, \mathcal{R}|_{\mathcal{V}-\{X\}})$ (this notation just says that X is removed from the set of endogenous variables and $\mathcal{R}$ is restricted so that its domain is $\mathcal{V} - \{X\}$ rather than all of $\mathcal{V}$) and $\mathcal{F}^{X=x}$ associates with each variable $Y \in \mathcal{V} - \{X\}$ the equation $F_Y^{X=x}$ which is obtained from F_Y by setting X to x. Thus, if M is the disjunctive causal model for the forest-fire example, then $M_{M=0}$, the model where the arsonist does not drop the match, has endogenous variables L and FF, where the equation for L is just as in M, and $FF = L$. If M is the conjunctive model, then equation for FF becomes instead $FF = 0$.

In this paper, following HP, I restrict to *acyclic* causal models, where causal influence can be represented by an acyclic Bayesian network. That is, there is no cycle $X_1, \ldots, X_n, X_1$ of endogenous variables where the value of X_{i+1} (as given by $F_{X_{i+1}}$) depends on the value of X_i, for $1 = 1, \ldots, n-1$, and the value of X_1 depends on the value of X_n. If M is an acyclic causal model, then given a *context*, that is, a setting $\vec{u}$ for the exogenous variables in $\mathcal{U}$, there is a unique solution for all the equations.

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

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

According to the definition of causality 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 forest 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, and hence the forest burning.

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

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

[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 I am trying to make here.

decisions often lie at the heart of determining actual causality in the real world. Disagreements about causality relationships often boil down to disagreements about the causal model. While the formalism presented here does not provide techniques to settle disputes about which causal model is the right one, at least it provides tools for carefully describing the differences between causal models, so that it should lead to more informed and principled decisions about those choices.

3 A Formal Definition of Actual Cause

3.1 A language for describing causes

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

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

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

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

For example, if M is the disjunctive causal model for the forest fire, and u is the context where there is lightning and the arsonist drops the lit match, then $(M, u) \models [M = 0](FF = 1)$, since even if the arsonist is somehow prevented from dropping the match, the forest burns (thanks to the lightning); similarly, $(M, u) \models [L = 0](FF = 1)$. However, $(M, u) \models [L = 0; M = 0](FF = 0)$: if arsonist does not drop the lit match and the lightning does not strike, then the forest does not burn.

3.2 A preliminary definition of causality

The HP definition of causality, like many others, is based on counterfactuals. The idea is that A is a cause of B if, if A hadn't occurred (although it did), then B would not have occurred. This idea goes back to at least Hume [1748, Section VIII], who said:

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

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

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

A more subtle problem is what philosophers have called *preemption*, where there are two potential causes of an event, one of which preempts the other. Preemption is illustrated by the following story taken from [Hall 2004]:

> Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle had 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. However, it does not satisfy the naive counterfactual definition either; if Suzy hadn't thrown, then Billy's throw would have shattered the bottle.

The HP definition deals with the first problem by defining causality as counterfactual dependency *under certain contingencies*. In the forest fire example, the forest fire does counterfactually depend on the lightning under the contingency that the arsonist does not drop the match; similarly, the forest fire depends oounterfactually on the arsonist's match under the contingency that the lightning does not strike. Clearly we need to be a little careful here to limit the contingencies that can be considered. We do not want to make Billy's throw the cause of the bottle shattering by considering the contingency that Suzy does not throw. The reason that we consider Suzy's throw to be the cause and Billy's throw not to be the cause is that Suzy's rock hit the bottle, while Billy's did not. Somehow the definition must capture this obvious intuition.

With this background, I now give the preliminary version of the HP definition of causality. Although the definition is labeled "preliminary", it is quite close to the final definition, which is given in Section 4. As I pointed out in the introduction, the definition is relative to a causal model (and a context); A may be a cause of B in one causal model but not in another. The definition consists of three clauses. The first and third are quite simple; all the work is going on in the second clause.

The types of events that the HP definition allows as actual causes are ones of the form $X_1 = x_1 \wedge \ldots \wedge X_k = x_k$—that is, conjunctions of primitive events; this is often abbreviated as $\vec{X} = \vec{x}$. The events that can be caused are arbitrary

Boolean combinations of primitive events. The definition does not allow statements of the form "A or A' is a cause of B," although this could be treated as being equivalent to "either A is a cause of B or A' is a cause of B". On the other hand, statements such as "A is a cause of B or B'" are allowed; as we shall see, this is not equivalent to "either A is a cause of B or A is a cause of B'".

Definition 3.1: (Actual cause; preliminary version) [Halpern and Pearl 2005] $\vec{X} = \vec{x}$ is an *actual cause of* φ *in* $(M, \vec{u})$ if the following three conditions hold:

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

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

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

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

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

$\vec{W}$, $\vec{w}$, and $\vec{x}'$ are said to be *witnesses* to the fact that $\vec{X} = \vec{x}$ is a cause of φ.

AC1 just says that $\vec{X} = \vec{x}$ cannot be considered a cause of φ unless both $\vec{X} = \vec{x}$ and φ actually happen. AC3 is a minimality condition, which ensures that only those elements of the conjunction $\vec{X} = \vec{x}$ that are essential for changing φ in AC2(a) are Clearly, all the "action" in the definition occurs in AC2. We can think of the variables in $\vec{Z}$ as making up the "causal path" from $\vec{X}$ to φ. Intuitively, changing the value of some variable in X results in changing the value(s) of some variable(s) in $\vec{Z}$, which results in the values of some other variable(s) in $\vec{Z}$ being changed, which finally results in the value of φ changing. The remaining endogenous variables, the ones in $\vec{W}$, are off to the side, so to speak, but may still have an indirect effect on what happens. AC2(a) is essentially the standard counterfactual definition of causality, but with a twist. If we want to show that $\vec{X} = \vec{x}$ is a cause of φ, we must show (in part) that if $\vec{X}$ had a different value, then so too would φ. However, this effect of the value of $\vec{X}$ on the value of φ may not hold in the actual context; the value of $\vec{W}$ may have to be different to allow this effect to manifest itself. For example, consider the context where both the lightning strikes and the arsonist drops a match in the disjunctive model of the forest fire. Stopping the arsonist from dropping the match will not prevent the forest fire. The counterfactual effect of the arsonist on the forest fire manifests itself only in a situation where the lightning does not strike (i.e., where L is set to 0). AC2(a) is what allows us to call both the lightning and the arsonist causes of the forest fire. Essentially, it ensures that $\vec{X}$ alone suffices to bring about the change from φ to $\neg\varphi$; setting $\vec{W}$ to $\vec{w}$ merely eliminates possibly spurious side effects that may mask the effect of changing the value of $\vec{X}$. Moreover, although the values of variables on the causal path (i.e., the variables $\vec{Z}$) may be perturbed by the change to $\vec{W}$, this perturbation has no impact on the value of φ. If $(M, \vec{u}) \models \vec{Z} = \vec{z}^*$, then $\vec{z}^*$ is the value of the variable Z in the context $\vec{u}$. We capture the fact that the perturbation has no impact on the value of φ by saying that if some variables Z on the causal path were set to their original values in the context $\vec{u}$, φ would still be true, as long as $\vec{X} = \vec{x}$.

To give some intuition for this definition, I consider three examples that will be relevant later in the paper.

Example 3.2: Can *not* performing an action be (part of) a cause? Consider the following story, also taken from (an early version of) [Hall 2004]: Suppose that Billy is hospitalized with a mild illness on Monday; he is treated and recovers. In the obvious causal model, the doctor's treatment is a cause of Billy's recovery. Moreover, if the doctor does *not* treat Billy on Monday, then the doctor's omission to treat Billy is a cause of Billy's being sick on Tuesday. But now suppose there are 100 doctors in the hospital. Although only doctor 1 is assigned to Billy (and he forgot to give medication), in principle, any of the other 99 doctors could have given Billy his medication. Is the nontreatment by doctors 2–100 also a cause of Billy's being sick on Tuesday? Of course, if we do not have variables in the model corresponding to the other doctors' treatment, or treat these variables as exogenous, then there is no problem. But if we have endogenous variables corresponding to the other doctors (for example, if we want to also consider other patients, who are being treated by these other doctors), then the other doctors' nontreatment is a cause, which seems inappropriate. I return to this issue in the next section.

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:

- T 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. I omit the obvious structural equations corresponding to this discussion.

In this causal model, it is true that $T = 1$ 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. $T = 1$ 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, $T = 1$ is *not* a cause of Billy's being alive. It fails condition AC2(a): setting $T = 0$ 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 T is, while if $TT = 1$, then Billy is dead when T 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 $T = 1$ is a cause of $TT = 0$ and $TT = 0$ is a cause of $BMC = 0 \vee BMC = 1 \vee BMC = 2$, $T = 1$ is not a cause of $BMC = 0 \vee BMC = 1 \vee BMC = 2$. Nor is causality closed under *right weakening*: $T = 1$ is a cause of $BMC = 0$, which logically implies $BMC = 0 \vee BMC = 1 \vee BMC = 2$, which is not caused by $T = 1$.

This distinguishes the HP definition from that of Lewis [2000], which builds in transitivity and implicitly assumes right weakening. ∎

The version of AC2(b) used here is taken from [Halpern and Pearl 2005], and differs from the version given in the conference version of that paper [Halpern and Pearl 2001]. In the current version, AC2(b) is required to hold for all subsets $\vec{W}'$ of $\vec{W}$; in the original definition, it was required to hold only for $\vec{W}$. The following example, due to Hopkins and Pearl [2003], illustrates why the change was made.

Example 3.3: 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 $D = 1$ in context u; given that B did not shoot (i.e., given that $B = 0$), A's loading the gun should not count as a cause. The obvious way to attempt to show that $A = 1$ is a cause is to take $\vec{W} = \{B, C\}$ and consider the contingency where $B = 1$ and $C = 0$. It is easy to check that AC2(a) holds for this contingency; moreover, $(M, u) \models [A = 1, B = 1, C = 0](D = 1)$. However, $(M, u) \models [A = 1, C = 0](D = 0)$. Thus, AC2(b) is not satisfied for the subset $\{C\}$ of W, so $A = 1$ is not a cause of $D = 1$. However, had we required AC2(b) to hold only for $\vec{W}$ rather than all subsets $\vec{W}'$ of $\vec{W}$, then $A = 1$ would have been a cause. ∎

While the change in AC2(b) has the advantage of being able to deal with Example 3.3 (indeed, it deals with the whole class of examples given by Hopkins and Pearl of which this is an instance), it has a nontrivial side effect. For the original definition, it was shown that the minimality condition AC3 guarantees that causes are always single conjuncts [Eiter and Lukasiewicz 2002; Hopkins 2001]. It was claimed in [Halpern and Pearl 2005] that the result is still true for the modified definition, but, as I now show, this is not the case.

Example 3.4: A and B both vote for a candidate. B's vote is recorded in two optical scanners (C_1 and C_2). If A votes for the candidate, then she wins; if B votes for the candidate and his vote is correctly recorded in the optical scanners, then the candidate wins. Unfortunately, A also has access to the scanners, so she will set them to read 0 if she does not vote for the candidate. In the actual context $\vec{u}$, both A and B vote for the candidate. The following structural equations characterize C and *WIN*: $C_i = \min(A, B)$, $i = 1, 2$, and $\mathit{WIN} = 1$ iff $A = 1$ or $C_1 = C_2 = 1$. I claim that $C_1 = 1 \wedge C_2 = 1$ is a cause of $\mathit{WIN} = 1$, but neither $C_1 = 1$ nor $C_2 = 1$ is a cause. To see that $C_1 = 1 \wedge C_2 = 1$ is a cause, first observe that AC1 clearly holds. For AC2, let $\vec{W} = \{A\}$ (so $\vec{Z} = \{B, C_1, C_2, \mathit{WIN}\}$) and take $w = 0$ (so we are considering the contingency where $A = 0$). Clearly, $(M, \vec{u}) \models [C_1 = 0, C_2 = 0, A = 0](\mathit{WIN} = 0)$ and $(M, \vec{u}) \models [C_1 = 1, C_2 = 1, A = a](\mathit{WIN} = 1)$, for both $a = 0$ and $a = 1$, so AC2 holds. To show that AC3 holds, I must show that neither $C_1 = 1$ nor $C_2 = 1$ is a cause of $\mathit{WIN} = 1$. The argument is the same for both $C_1 = 1$ and $C_2 = 1$, so I just show that $C_1 = 1$ is not a cause. To see this, note that if $C_1 = 1$ is a cause with $\vec{W}$, $\vec{w}$, and $\vec{x}'$ as witnesses, then $\vec{W}$ must contain A and $\vec{w}$ must be such that $A = 0$. But since $(M, u) \models [C_1 = 1, A = 0](\mathit{WIN} = 0)$, AC2(b) is violated no matter whether C_2 is in $\vec{Z}$ or in $\vec{W}$. ∎

Although Example 3.4 shows that causes are not always single conjuncts, they often are. Indeed, it is not hard to show that in all the standard examples considered in the philosophy and legal literature (in particular, in all the examples considered in HP), they are. The following result give some intuition as to why. Further intuition is given by the results of Section 5. Notice that in Example 3.4, A affects both C_1 and C_2. As the following result shows, we do not have conjunctive causes if the potential causes cannot be affected by other variables.

Say that $\vec{X} = \vec{x}$ is a *weak cause of* φ *under the contingency* $\vec{W} = \vec{w}$ *in* $(M, \vec{u})$ if AC1 and AC2 hold under the contingency $\vec{W} = \vec{w}$, but AC3 does not necessarily hold.

Proposition 3.5: *If $\vec{X} = \vec{x}$ is a weak cause of φ in $(M, \vec{u})$ with $\vec{W}$, $\vec{w}$, and $\vec{x}'$ as witnesses, $|\vec{X}| > 1$, and each variable X_i in $\vec{X}$ is independent of all the variables in $\mathcal{V} - \vec{X}$ in $\vec{u}$ (that is, if $\vec{Y} \subseteq \mathcal{V} - \vec{X}$, then for each setting $\vec{y}$ of $\vec{Y}$, we have $(M, \vec{u}) \models \vec{X} = \vec{x}$ iff $(M, \vec{u}) \models [\vec{Y} = \vec{y}](\vec{X} = \vec{x})$), then $\vec{X} = \vec{x}$ is not a cause of φ in $(M, \vec{u})$.*

In the examples in [Halpern and Pearl 2005] (and elsewhere in the literature), the variables that are potential causes are typically independent of all other variables, so in these causes are in fact single conjuncts.

4 Dealing with normality and typicality

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

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

Example 4.1 illustrates an even deeper problem with Definition 3.1. The structural equations for Example 4.1 are *isomorphic* to those in the forest-fire example, provided that we interpret the variables appropriately. Specifically, take the endogenous variables in Example 4.1 to be A (for "assassin does not put in poison"), B (for "bodyguard puts in antidote"), and VS (for "victim survives"). Then A, B, and VS satisfy exactly the same equations as L, M, and FF, respectively. In the context where there is lightning and the arsonists drops a lit match, both the the lightning and the match are causes of the forest fire, which seems reasonable. But here it does not seem reasonable that Bodyguard's putting in the antidote is a cause. Nevertheless, any definition that just depends on the structural equations is bound to give the same answers in these two examples. (An example illustrating the same phenomenon is given by Hall [2007].) This suggests that there must be more to causality than just the structural equations. And, indeed, the final HP definition of causality allows certain contingencies to be labeled as "unreasonable" or "too farfetched"; these contingencies are then not considered in AC2(a) or AC2(b). Unfortunately, it is not always clear what makes a contingency unreasonable. Moreover, this approach will not work to deal with Example 3.2.

In this example, we clearly want to consider as reasonable the contingency where no doctor is assigned to Billy and Billy is not treated (and thus is sick on Tuesday). We should also consider as reasonable the contingency where doctor 1 is assigned to Billy and treats him (otherwise we cannot say that doctor 1 is the cause of Billy being sick if he is assigned to Billy and does not treat him). What about the contingency where doctor $i > 1$ is assigned to treat Billy and does so? It seems just as reasonable as the one where doctor 1 is assigned to treat Billy and does so. Indeed, if we do not call it reasonable, then we will not be able to say that doctor i is a cause of Billy's sickness in the context where doctor i assigned to treat Billy and does not. On the other hand, if we call it reasonable, then if doctor 1 is assigned to treat Billy and does not, then doctor $i > 1$ not treating Billy will also be a cause of Billy's sickness. To deal with this, what is reasonable will have to depend on the context; in the context where doctor 1 is assigned to treat Billy, it should not be considered reasonable that doctor $i > 1$ is assigned to treat Billy.

As suggested in the introduction, the solution involves assuming that an agent has, in addition to a theory of causality (as modeled by the structural equations), a theory of "normality" or "typicality". This theory would include statements like "typically, people do not put poison in coffee" and "typically doctors do not treat patients to whom they are not assigned". There are many ways of giving semantics to such typicality statements, including *preferential structures* [Kraus, Lehmann, and Magidor 1990; Shoham 1987], *ϵ-semantics* [Adams 1975; Geffner 1992; Pearl 1989], and *possibilistic structures* [Dubois and Prade 1991], and ranking functions [Goldszmidt and Pearl 1992; Spohn 1988]. For definiteness, I use the last approach here (although it would be possible to use any of the other approaches as well).

Take a *world* to be a complete description of the values of all the random variables. I assume that each world has associated with it a *rank*, which is just a natural number or ∞. Intuitively, the higher the rank, the less likely the world. A world with a rank of 0 is reasonably likely, one with a rank of 1 is somewhat likely, one with a rank of 2 is quite unlikely, and so on. Given a ranking on worlds, the statement "if p then typically q" is true if in all the worlds of least rank where p is true, q is also true. Thus, in one model where people do not typically put either poison or antidote in coffee, the worlds where neither poison nor antidote is put in the coffee have rank 0, worlds where either poison or antidote is put in the coffee have rank 1, and worlds where both poison and antidote are put in the coffee have rank 2.

Take an *extended causal model* to be a tuple $M = (\mathcal{S}, \mathcal{F}, \kappa)$, where $(\mathcal{S}, \mathcal{F})$ is a causal model, and κ is a *ranking function* that associates with each world a rank. In an acyclic extended causal model, a context $\vec{u}$ determines a world denoted $s_{\vec{u}}$. $\vec{X} = \vec{x}$ is a *cause of φ in an extended model M and context $\vec{u}$* if $\vec{X} = \vec{x}$ is a cause of φ according to Definition 3.1, except that in AC2(a), there must be a world s such that $\kappa(s) \leq \kappa(s_{\vec{u}})$ and $\vec{X} = \vec{x}' \wedge \vec{W} = \vec{w}$ is true at s. This can be viewed as a formalization of Kahnemann and Miller's observation that we tend to alter the exceptional than the routine aspects of a world; we consider only alterations that hold in a world that is no more exceptional than the actual world.[3] (The idea of extending causal models with a ranking function already appears in [Halpern and Pearl 2001], but it was not used to capture statements about typicality as suggested here. Rather, it was used to talk about $\vec{X} = \vec{x}$ being a cause of φ *at rank k*, where k is the lowest rank of the world that shows that $\vec{X} = \vec{x}$ is a cause. The idea was dropped in the journal version of the paper.)

This definition deals well with all the problematic examples in the literature. Consider Example 4.1. Using the rank-

[3] I originally considered requiring that $\kappa(s) < \kappa(s_{\vec{u}})$, so that you move to a strictly more normal world, but this seems too strong a requirement. For example, suppose that A wins an election over B by a vote of 6–5. We would like to say that each voter for A is a cause of A's winning. But if we view all voting patterns as equally normal, then no voter is a cause of A's winning, because no contingency is more normal than any other.

ing described above, Bodyguard is not a cause of Victim's survival because the world that would need to be considered in AC2(a), where Assassin poison the coffee, is less normal than the actual world, where he does not. It also deals well with Example 3.2. Suppose that in fact the hospital has 100 doctors and there are variables $A_1, \ldots, A_{100}$ and $T_1, \ldots, T_{100}$ in the causal model, where $A_i = 1$ if doctor i is assigned to treat Billy and $A_i = 0$ if he is not, and $T_i = 1$ if doctor i actually treats Billy on Monday, and $T_i = 0$ if he does not. Doctor 1 is assigned to treat Billy; the others are not. However, in fact, no doctor treats Billy. Further assume that typically, doctors do not treat patients (that is, a random doctor does not typically treat a random patient), and if doctor i is assigned to Billy, then typically doctor i treats Billy. We can capture this in an extended causal model where the world where no doctor is assigned to Billy and no doctor treats him has rank 0; the 100 worlds where exactly one doctor is assigned to Billy, and that doctor treats him, have rank 1; the 100 worlds where exactly one doctor is assigned to Billy and no one treats him have rank 2; and the 100×99 worlds where exactly one doctor is assigned to Billy but some doctor treats him have rank 3. (The ranking given to other worlds is irrelevant.) In this extended model, in the context where doctor i is assigned to Billy but no one treats him, i is the cause of Billy's sickness (the world where i treats Billy has lower rank than the world where i is assigned to Billy but no one treats him), but no other doctor is a cause of Billy's sickness. Moreover, in the context where i is assigned to Billy and treats him, then i is the cause of Billy's recovery (for AC2(a), consider the world where no doctor is assigned to Billy and none treat him).

I consider one more example here, due to Hitchcock [2007], that illustrates the interplay between normality and causality.

Example 4.2: Assistant Bodyguard puts a harmless antidote in Victim's coffee. Buddy then poisons the coffee, using a type of poison that is normally lethal, but is countered by the antidote. Buddy would not have poisoned the coffee if Assistant had not administered the antidote first. (Buddy and Assistant do not really want to harm Victim. They just want to help Assistant get a promotion by making it look like he foiled an assassination attempt.) Victim drinks the coffee and survives. ∎

Is Assistant's adding the antidote a cause of Victim's survival? Using the preliminary HP definition, it is; if Assistant does not add the antidote, Victim survives. However, using an extended causal model with the normality assumptions implied by the story, it is not. Specifically, suppose we assume that if Assistant does not add the antidote, then Buddy does not normally add poison. (Buddy, after all, is normally a law-abiding citizen.) In the corresponding extended causal model, the world where Buddy poisons the coffee and Assistant does not add the Antidote has a higher rank (i.e., is less normal than) the world where Buddy poisons the coffee and Assistant adds the antidote. This is all we need to know about the ranking function to conclude that adding the antidote is not a cause. By way of contrast, if Buddy were a more typical assassin, with reasonable normality assumptions, the world where he puts in the poison and Assistant puts in the antidote would be less normal than then one Buddy puts in the poison and Assistant does not put in the antidote, so Assistant would be a cause of Victim being a alive.

Interestingly, Hitchcock captures this story using structural equations that also make Assistant putting in the antidote a *cause* of Buddy putting in the poison. This is the device used to distinguish this situation from one where Buddy is actually means Victim to die (in which case Buddy would presumably have put in the poison even if Assistant had not added the antidote). However, it is not clear that people would agree that Assistant putting in the antidote really *caused* Buddy to add the poison; rather, it set up a circumstance where Buddy was willing to put it in. I would argue that this is better captured by using the normality statement "If Assistant does not put in the antidote, then Buddy does not normally add poison." As this example shows, there is a nontrivial interplay between statements of causality and statements of normality.

I leave it to the reader to check that reasonable assumptions about typicality can also be used to deal with the other problematic examples for the HP definition that have been pointed out in the literature, such as Larry the Loanshark [Halpern and Pearl 2005, Example 5.2] and Hall's [2007] watching police example. (The family sleeps peacefully through the night. Are the watching police a cause? After all, if there had been thieves, the police would have nabbed them, and without the police, the family's peace would have been disturbed.)

This is not the first attempt to modify structural equations to deal with defaults; Hitchcock [2007] and Hall [2007] also consider this issue. Neither adds any extra machinery such as ranking functions, but both assume that there is an implicitly understood notion of normality. Roughly speaking, Hitchcock [2007] can be understood as giving constraints on models that guarantee that the answer obtained using the preliminary HP definition agrees with the answer obtained using the definition in extended causal models. I do not compare my suggestion to that of Hall [2007], since, as Hitchcock [2008] points out, there are a number of serious problems with Hall's approach. It is worth noting that both Hall and Hitchcock assume that a variable has a "normal" or "default" setting; any other setting is abnormal. However, it is easy to construct examples where what counts as normal depends on the context. For example, it is normal for doctor i to treat Billy if i is assigned to Billy; otherwise it is not.

5 The NESS approach

In this section I provide a sufficient condition to guarantee that a single conjunct is a cause. Doing so has the added benefit of providing a careful comparison of the NESS test and the HP approach. Wright does not provide a mathematical formalization of the NESS test; what I give here is my understanding of it.

A is a cause of B according to the NESS test if there exists a set $\mathbf{S} = \{A_1, \ldots, A_k\}$ of events, each of which actually occurred, where $A = A_1$, $\mathbf{S}$ is sufficient for for B, and $\mathbf{S} - \{A_1\}$ is not sufficient for B. Thus, A is an element of a sufficient condition for B, namely $\mathbf{S}$, and is a necessary element of that set, because any subset of $\{A_1, \ldots, A_k\}$ that does not include A is not sufficient for B.[4]

The NESS test, as stated, seems intuitive and simple. Moreover, it deals well with many examples. However, although the NESS test looks quite formal, it lacks a definition of what it means for a set $\mathbf{S}$ of events to be *sufficient* for B to occur. As I now show, such a definition is sorely needed.

Example 5.1: Consider Wright's example of Victoria's poisoning from the introduction. First, suppose that Victoria drinks a cup of tea poisoned by Paula, and then dies. It seems clear that Paula poisoning the tea caused Victoria's death. Let $\mathbf{S}$ consist of two events:

- A_1, Paula poisoned the tea; and
- A_2, Victoria drank the tea.

Given our understanding of the world, it seems reasonable to say that the A_1 and A_2 are sufficient for Victoria's death, but removing A_1 results in a set that is insufficient.

But now suppose that Sharon shoots Victoria just after she drinks the tea (call this event A_3), and she dies instantaneously from the shot (before the poison can take effect). In this case, we would want to say that A_3 is the cause of Victoria's death, not A_2. Nevertheless, it would seem that the same argument that makes Paula's poisoning a cause without Sharon's shot would still make Paula's poisoning a cause even without Sharon's shot. The set $\{A_1, A_2\}$ still seems sufficient for Victoria's death, while $\{A_2\}$ is not.

Wright [1985] observes the poisoned tea would be a cause of Victoria's death only if Victoria "drank the tea and *was alive when the poison took effect*". Wright seems to be arguing that $\{A_1, A_2\}$ is in fact *not* sufficient for Victoria's death. We need A_3: Victoria was alive when the poison took effect. While I agree that the fact that Victoria was alive when the poison took place is critical for causality, I do not see how it helps in the NESS test, under what seems to me the most obvious definitions of "sufficient". I would argue that $\{A_1, A_2\}$ *is* in fact just as sufficient for death as $\{A_1, A_2, A_3\}$. For suppose that A_1 and A_2 hold. Either Victoria was alive when the poison took effect, or she was not. In the either case, she dies. In the former case, it is due to the poison; in the latter case, it is not.

But it gets worse. While I would argue that $\{A_1, A_2\}$ is indeed just as sufficient for death as $\{A_1, A_2, A_3\}$, it is not clear that $\{A_1, A_2\}$ is in fact sufficient. Suppose, for example, that some people are naturally immune to the poison that Paula used, and do not die from it. Victoria is not immune. But then it seems that we need to add a condition A_4 saying that Victoria is not immune from the poison to get a set sufficient to cause Victoria's death. And why should it stop there? Suppose that the poison has an antidote that, if administered within five minutes of the poison taking effect, will prevent death. Unfortunately, the antidote was not administered to Victoria, but do we have to add this condition to $\mathbf{S}$ to get a sufficient set for Victoria's death? Where does it stop? ∎

The NESS definition is also unclear as to which events can go in $\mathbf{S}$. The problem is illustrated in the next example.

Example 5.2: Wright [2001] considers an example where defendant 1 discharged 15 units of effluent, while defendant 2 discharged 13 units. Suppose that 14 units of effluent are sufficient for injury. It seems clear that defendant 1's discharge is a cause of injury; if he hadn't discharged any effluent, then there would have been no injury. What about defendant 2's discharge? In the HP approach, whether it is a cause depends on the random variables considered and their possible values. Suppose that D_i is a random variable representing defendant i's discharge, for $i = 1, 2$. If D_1 can only take values 0 or 15 (i.e., if defendant 1 discharges either nothing or all 15 units), then defendant 2's discharge is not a cause. But if D_1 can take, for example, every integer value between 0 and 15, then $D_2 = 13$ is a cause (under the contingency that $D_1 = 4$, for example).

Intuitively, the decision as to whether the causal model should include 4 as a possible value of D_1 or have 0 and 15 as the only possible values of D_1 should depend on the options available to defendant 1. If all he can do is to press a switch that determines whether or not there is effluent (so that pressing the switch results in D_1 being 15, and not pressing it result in D_1 being 0) then it seems reasonable to take 0 and 15 as the only values. On the other hand, if the defendant can control the amount of effluent, then taking the range of values to include every number between 0 and 15 seems more reasonable.

Perhaps not surprisingly, this issue is relevant to the NESS test as well, for the same reason. If the only possible values of D_1 are 0 or 15, then there is no set $\mathbf{S}$ including $D_2 = 13$ that is sufficient for the injury such that $D_2 = 13$ is necessary. On the other hand, if $D_1 = 4$ is a possible event, then there is such a set. ∎

The problem raised by Example 5.2, that of which events can go into $\mathbf{S}$, is easy to deal with, by simply making the set of variables that can go into $\mathbf{S}$ explicit. Of course, as the example suggests, the choice of events will have an impact on what counts as a cause, but that is arguably appropriate. Recall that causal models deal with this issue by making explicit the signature, that is, the set of variables and their possible values. This gives us a set of primitive events of the form $X = x$. More complicated events can be formed as Boolean combinations of primitive events, but it may also be reasonable to restrict $\mathbf{S}$ to consisting of only primitive events.

The problem raised by Example 5.1, that of defining sufficient cause, seems more serious. I believe that a formal definition will require some of the machinery of causal models, including structural equations. (This point echoes criticisms of NESS and related approaches by Pearl [2000, pp. 314–315].) I now sketch an approach to defining sufficiency that delivers reasonable answers in many cases of interest and, indeed, often agrees with the HP definition.[5]

[4]The NESS test is much in the spirit of Mackie's INUS test [Mackie 1965], according to which A is a cause of B if A is an insufficient but necessary part of a condition which is unnecessary but sufficient for B. However, a comparison of the two approaches is beyond the scope of this paper.

[5]Interestingly, Baldwin and Neufeld [2003] claimed that the

Fix a causal model M. Recall that a primitive event has the form $X = x$; a set of primitive events is *consistent* if it does not contain both $X = x$ and $X = x'$ for some random variable X and $x \neq x'$. If $\mathbf{S} = \{X_1 = x_1, \ldots, X_k = x_k\}$ is a consistent set of primitive events, then $\mathbf{S}$ is *sufficient* for φ relative to causal model M if $M \models [\mathbf{S}]\varphi$, where $[\mathbf{S}]\varphi$ is an abbreviation for $[X_1 = x_1; \ldots; X_k = x_k]\varphi$. Roughly speaking, the idea is to formalize the NESS test by taking $X = x$ to be a cause of φ if there is a a set $\mathbf{S}$ including $X = x$ that is sufficient for φ, while $\mathbf{S} - \{X = x\}$ is not. Example 5.1 already shows that this will not work. If CP is a random variable that takes on value 1 if Paula poisoned the tea and 0 otherwise, then it is not hard to show that in the obvious causal model, $CP = 1$ is sufficient for $PD = 1$ (Victoria dies), even if Sharon shoots Victoria. To deal with this problem, we must strengthen the notion of sufficiency to capture some of the intuitions behind AC2(b).

Say that $\mathbf{S}$ is *strongly sufficient for φ in $(M, \vec{u})$* if $\mathbf{S} \cup \mathbf{S}'$ is sufficient for φ in M for all sets $\mathbf{S}'$ consisting of primitive events $Z = z$ such that $(M, \vec{u}) \models Z = z$. Intuitively, $\mathbf{S}$ is strongly sufficient for φ in $(M, \vec{u})$ if $\mathbf{S}$ remains sufficient for φ even when additional events, which happen to be true in $(M, \vec{u})$, are added to it. As I now show, although $CP = 1$ is sufficient for $PD = 1$, it is not strongly sufficient, provided that the language includes enough events.

As already shown by HP, in order to get the "right" answer for causality in the presence of preemption (here, the shot preempts the poison), there must be a variable in the language that takes on different values depending on which of the two potential causes is the actual cause. In this case, we need a variable that takes on different values depending on whether Sharon shot. Suppose that it would take Victoria t units of time after the poison is administered to die; let DAP be the variable that has value 1 if Victoria dies t units of time after the poison is administered and is alive before that, and has value 0 otherwise. Note that $DAP = 0$ if Victoria is already dead before the poison takes effect. In particular, if Sharon shoots Victoria before the poison takes effect, then $DAP = 0$. Then although $CP = 1$ is sufficient for $PD = 1$, it is not strongly sufficient for $PD = 1$ in the context $\vec{u}'$ where Sharon shoots, since $(M, \vec{u}) \models DAP = 0$, and $M \models [CP = 1; DAP = 0](PD \neq 1)$.

The following definition is my attempt at formalizing the NESS condition, using the ideas above.

Definition 5.3: $\vec{X} = \vec{x}$ is a *cause of φ in $(M, \vec{u})$ according to the causal NESS test* if there exists a set $\mathbf{S}$ of primitive events containing $\vec{X} = \vec{x}$ such that the following properties hold:

NT1. $(M, \vec{u}) \models \mathbf{S}$; that is, $(M, \vec{u}) \models Y = y$ for all primitive events $Y = y$ in $\mathbf{S}$.

NT2. $\mathbf{S}$ is strongly sufficient for φ in $(M, \vec{u})$.

NT3. $\mathbf{S} - \{\vec{X} = \vec{x}\}$ is not strongly sufficient for φ in $(M, \vec{u})$.

NT4. $\vec{X} = \vec{x}$ is minimal; no subset of $\vec{X}$ satisfies conditions NT1–3.[6]

$\mathbf{S}$ is said to be a *witness* for the fact that $\vec{X} = \vec{x}$ is a cause of φ according to the causal NESS test. ∎

Unlike the HP definition, causes according to the causal NESS test always consist of single conjuncts.

Theorem 5.4: *If $\{X_1 = x_1, \ldots, X_k = x_k\}$ is a cause of φ in M according to the causal NESS test, then $k = 1$.*

It is easy to check that in Example 3.4, both $C_1 = 1$ and $C_2 = 1$ are causes of $WIN = 1$ according to the causal NESS test, while (because of NT4) $C_1 = 1 \wedge C_2 = 1$ is not. On the other hand, Example 3.4 shows that neither $C_1 = 1$ nor $C_2 = 1$ is a cause according to the HP definition, while $C_1 \wedge C_2 = 1$ is. Thus, the two definitions are incomparable.

Nevertheless, the HP definition and the causal NESS test agree in many cases of interest (in particular, in all the examples in the HP paper). In light of Theorem 5.4, this explains in part why, in so many cases, causes are single conjuncts with the HP definition. In the rest of this section I give conditions under which the NESS test and the HP definition agree. Although they are complicated, they apply in all the standard examples in the literature.

I start with conditions that suffice to show that being a cause with according to the causal NESS test implies being a cause according to the HP definition.

Theorem 5.5: *Suppose that $X = x$ is a cause of φ in $(M, \vec{u})$ according to the causal NESS test with witness $\mathbf{S}$, and there exists a (possible empty) set $\vec{T}$ of variables not mentioned in φ or $\mathbf{S}$ and a context $\vec{u}'$ such that the following properties hold:*

SH1. *$\mathbf{S} - \{X = x\}$ is not a sufficient condition for φ in $(M, \vec{u}')$; that is, $(M, \vec{u}') \models [\mathbf{S} - \{X = x\}]\neg\varphi$.*[7]

SH2. *Each variable in $\vec{T}$ is independent of all other variables in contexts $\vec{u}$ and $\vec{u}'$; that is, for all variables $T \in \vec{T}$, if $\vec{W}$ consists of all endogenous variables other than T, then for all settings t of T and $\vec{w}$ of $\vec{W}$, we have $(M, \vec{u}) \models T = t$ iff $(M, \vec{u}) \models [\vec{W} = \vec{w}](T = t)$, and similarly for context $\vec{u}'$.*

SH3. *φ is determined by $\vec{T}$ and X in contexts $\vec{u}$ and $\vec{u}'$; that is, for all $\vec{t}$, $\vec{T}'$ disjoint from $\vec{T}$ and X, x', and $\vec{t}'$, we have $(M, \vec{u}') \models [\vec{T} = \vec{t}, \vec{T}' = \vec{t}', X = x']\varphi$ iff $(M, \vec{u}) \models [\vec{T} = \vec{t}, \vec{T}' = \vec{t}', X = x']\varphi$.*

SH4. *In context $\vec{u}$, $\mathbf{S} - \{X = x\}$ depends only on $X = x$ in $\vec{u}$; that is, for all $\vec{T}'$ disjoint from $\mathbf{S}$ and $\vec{t}'$, we have $(M, \vec{u}) \models [\vec{X} = x, \vec{T}' = \vec{t}']\mathbf{S}$.*

NESS test could be formalized using causal models, but did not actually show how, beyond describing some examples. In a later paper [Baldwin and Neufeld 2004], they seem to retract the claim that the NESS test can be formalized using causal models.

[6] This definition does not take into account defaults. It can be extended to take defaults into account by requiring that if $\vec{u}'$ is the context showing that $\mathbf{S} - \{X = x\}$ is not strongly sufficient for φ in NT2, then $\kappa(s_{\vec{u}'}) \leq \kappa(s_{\vec{u}})$. For ease of exposition, I ignore this issue here.

[7] Since $\mathbf{S}$ is a witness to the fact that $X = x$ is a cause of φ in $(M, \vec{u})$, $\mathbf{S} - \{X = x\}$ is not a strongly sufficient cause for φ with respect to $(M, \vec{u})$. SH1 requires something different: that $\mathbf{S} - \{X = x\}$ not be a sufficient cause for φ in $(M, \vec{u}')$.

Then $X = x$ is a cause of φ in $(M, \vec{u})$ according to the HP definition.

Getting conditions sufficient for causality according to the HP definition to imply causality according to the NESS test is not so easy. The problem is the requirement in the NESS definition that there be a witness $\mathbf{S}$ such that $(M, \vec{u}') \models [\mathbf{S}]\varphi$ in *all* contexts $\vec{u}'$ is very strong, indeed, arguably too strong. For example, consider a vote that might be called off if the weather is bad, where the weather is part of the context. Thus, in a context where the weather is bad, there is no winner, even if some votes have been cast. In the actual context, the weather is fine and A votes for Mr. B, who wins the election. A's vote is a cause of Mr. B's victory in this context, according to the HP definition, but not according to the NESS test, since there is no set $\mathbf{S}$ that includes A sufficient to make Mr. B win in all contexts; indeed, there is no cause for Mr. B's victory according to the NESS test (which arguably indicates a problem with the definition).

Since the HP definition just focuses on the actual context, there is no obvious way to conclude from $X = x$ being a cause of φ in context $\vec{u}$ a condition holds in all contexts. To deal with this, I weaken the NESS test so that it must hold only with respect to a set U of contexts. More precisely, say that $\mathbf{S}$ *is sufficient for φ with respect to U* if $(M, u) \models [\mathbf{S}]\varphi$ for all $u \in U$. We can then define what it means for $\mathbf{S}$ to be *strongly sufficient for φ in $(M, \vec{u})$ with respect to U* and for $\vec{X} = \vec{x}$ to be a *cause of φ in $(M, \vec{u})$ with respect to U* in the obvious way; in the latter case, we simply require take strong sufficiency in NT2 and NT3 to be with respect to U. It is easy to check that Theorem 5.4 holds (with no change in proof) for causality with respect to a set U of contexts; that is, even in this case, a cause must be a single conjunct.

Theorem 5.6: *Suppose that $X = x$ is a cause of φ in $(M, \vec{u})$ according to the HP definition, with $\vec{W}$, $\vec{w}$, and x' as witnesses. Suppose that there exists a subset $\vec{W}' \subseteq \vec{W}$ such that $(M, \vec{u}') \models \vec{W}' = \vec{w}$ (that is, the assignment $\vec{W}' = \vec{w}$ does not change the values of the variables in $\vec{W}'$ in context $(M, \vec{u})$) and a context $\vec{u}'$ such that the following conditions hold, where $\vec{W}'' = \vec{W} - \vec{W}'$:*

SN1. $(M, \vec{u}') \models [\vec{W}' = \vec{w}](X = x' \wedge \vec{W}'' = \vec{w})$.

SN2. *$\vec{W}''$ is independent of $\vec{Z}$ given $X = x$ and $\vec{W} = \vec{w}$ in $\vec{u}'$, so that if $\vec{Z}' \subseteq \vec{Z}$, then for all $\vec{z}'$, we have $(M, \vec{u}') \models [X = x, \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}'](\vec{W}'' = \vec{w})$.*

SN3. *φ is independent of $\vec{u}$ and $\vec{u}'$ conditional on X and $\vec{W} = \vec{w}$; that is if $\vec{Z}' \subseteq \vec{Z}$, then for all $\vec{z}'$ and x'', we have $(M, \vec{u}') \models [X = x'', \vec{W} = \vec{w}', \vec{Z} = \vec{z}']\varphi$ iff $(M, \vec{u}) \models [X = x'', \vec{W} = \vec{w}', \vec{Z} = \vec{z}']\varphi$.*

Then $X = x$ is a cause of φ in $(M, \vec{u})$ with respect to $\{\vec{u}, \vec{u}'\}$ according to the causal NESS test.

6 Discussion

It has long been recognized that normality is a key component of causal reasoning. Here I show how it can be incorporated into the HP framework in a straightforward way. The HP approach defines causality relative to a causal model. But we may be interested in whether a causal statement follows from some features of the structural equations and some default statements, without knowing the whole causal model. For example, in a scenario with many variables, it may be infeasible (or there might not be enough information) to provide all the structural equations and a complete ranking function. This suggests it may be of interest to find an appropriate logic for reasoning about actual causality. Axioms for causal reasoning (expressed in the language of this paper, using formulas of the form $[\vec{X} = \vec{x}]\varphi$, have already been given by Halpern [2000]; the KLM axioms [Kraus, Lehmann, and Magidor 1990] for reasoning about normality and defaults are well known. It would be of interest to put these axioms together, perhaps incorporating ideas from the causal NESS test, and adding some statements about (strong) sufficiency, to see if they lead to interesting conclusions about actual causality.

Acknowledgments: I thank Steve Sloman for pointing out [Kahneman and Miller 1986], Denis Hilton and Chris Hitchcock for intersting discussions on causality, and Judea Pearl and the anonymous KR reviewers for useful comments.

References

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

Baldwin, R. A. and E. Neufeld (2003). On the structure model interpretation of Wright's NESS test. In *Proc. AI 2003*, Lecture Notes in AI, Vol. 2671, pp. 9–23.

Baldwin, R. A. and E. Neufeld (2004). The structural model interpretation of the NESS test. In *Advances in Artificial Intelligence*, Lecture Notes in Computer Science, Vol. 3060, pp. 297–307.

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

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

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

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

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

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

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

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 Conf. on Uncertainty in Artificial Intelligence (UAI 2001)*, pp. 194–202.

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

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

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

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

Hitchcock, C. (2008). Structural equations and causation: six counterexamples. *Philosophical Studies*.

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

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

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

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

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

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 Conf. on Artificial Intelligence (IJCAI '95)*, pp. 1985–1991.

Mackie, J. (1965). Causes and conditions. *American Philosophical Quarterly 2/4*, 261–264.

Pearl, J. (1989). Probabilistic semantics for nonmonotonic reasoning: a survey. In *Proc. First International Conf. on Principles of Knowledge Representation and Reasoning (KR '89)*, pp. 505–516.

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

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

Sandewall, E. (1994). *Features and Fluents*, Vol. 1. Clarendon Press.

Shoham, Y. (1987). A semantical approach to nonmonotonic logics. In *Proc. 2nd IEEE Symposium on Logic in Computer Science*, pp. 275–279.

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

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

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

Wright, R. W. (2001). Once more into the bramble bush: Duty, causal contribution, and the extent of legal responsibility. *Vanderbilt Law Review 54*(3), 1071–1132.

A Appendix: Proofs

In this appendix, I prove the results stated in the text. For the reader's convenience, I repeat the statement of the results here.

Proposition 3.5: *If $\vec{X} = \vec{x}$ is a weak cause of φ in $(M, \vec{u})$ with $\vec{W}$, $\vec{w}$, and $\vec{x}'$ as witnesses, $|\vec{X}| > 1$, and each variable X_i in $\vec{X}$ is independent of all the variables in $\mathcal{V} - \vec{X}$ in $\vec{u}$ (that is, if $\vec{Y} \subseteq \mathcal{V} - \vec{X}$, then for each setting $\vec{y}$ of $\vec{Y}$, we have $(M, \vec{u}) \models \vec{X} = \vec{x}$ iff $(M, \vec{u}) \models [\vec{Y} = \vec{y}](\vec{X} = \vec{x})$)*
, then $\vec{X} = \vec{x}$ is not a cause of φ in $(M, \vec{u})$.

Proof: Suppose that the hypotheses of the proposition hold. First note that since $\vec{X} = \vec{x}$ is a weak cause of φ in $(M, \vec{u})$, by AC1, we must have $(M, \vec{u}) \models \vec{X} = \vec{x}$. Since each variable in $\vec{X}$ is independent of all the variables in $\mathcal{V} - \vec{X}$, for all $\vec{Y} \subseteq \mathcal{V} - \vec{X}$ and all settings $\vec{y}$ of the variables in $\vec{Y}$, we must have $(M, \vec{u}) \models [\vec{Y} = \vec{y}](\vec{X} = \vec{x})$. It follows that, for all formulas ψ, all subsets $\vec{X}'$ of $\vec{X}$, all subsets $\vec{Y}$ of $\mathcal{V} - \vec{X}$, and all settings $\vec{y}$ of $\vec{Y}$, we have

$$(M, \vec{u}) \models [\vec{Y} = \vec{y}]\psi \text{ iff } (M, \vec{u}) \models [\vec{X}' = \vec{x}, \vec{Y} = \vec{y}]\psi. \quad (1)$$

Next, observe that since the causal model is acyclic, there must be some variable in $\vec{X}$ that is independent of every other variable in $\vec{X}$. Without loss of generality, suppose that it is X_1. Thus, X_1 is independent of every variable in $\mathcal{V} - \{X_1\}$. Let $\vec{X}^- = \langle X_2, \ldots, X_k \rangle$. I show that either $X_1 = x_1$ or $\vec{X}^- = \vec{x}$ is a weak cause of φ, showing that $\vec{X} = \vec{x}$ is not a cause of φ, since it does not satisfy AC3.

First suppose that $x_1 = x_1'$. I show that then $\vec{X}^- = \vec{x}$ is a weak cause of φ, with $\vec{W}$, $\vec{w}$, and $\vec{x}'$ as witnesses. To see this, note that since $\vec{X} = \vec{x}$ is a weak cause of φ, with $\vec{W}$, $\vec{w}$, and $\vec{x}'$ as witnesses, by AC2(a), we have that $(M, \vec{u}) \models [\vec{X} = \vec{x}', \vec{W} = \vec{w}]\neg\varphi$. By the same arguments as used to derive (1), we have that $(M, \vec{u}) \models [\vec{X}^- = \vec{x}, \vec{W} = \vec{w}]\neg\varphi$. Thus, AC2(a) holds for $\vec{X}^- = \vec{x}$. By AC2(b), $(M, \vec{u}) \models [\vec{X} = \vec{x}, \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}^*]\varphi$ for all subsets $\vec{W}'$ of $\vec{W}$ and all subsets $\vec{Z}'$ of $\vec{Z}$. By (1), we have that $(M, \vec{u}) \models [\vec{X}^- = \vec{x}, \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}^*]\varphi$. Thus, AC2(b) holds for $\vec{X}^- = \vec{x}$, and $\vec{X}^- = \vec{x}$ is indeed a weak cause of φ.

Now suppose that $x_1 \neq x_1'$. If $\vec{X}^- = \vec{x}$ is a weak cause of φ with witnesses $\vec{W} \cup \{X_1\}$, $\vec{w} \cdot \langle x_1' \rangle$, and $\vec{x}$, then we are done. So suppose that $\vec{X}^- = \vec{x}$ is not a weak cause of φ with witnesses $\vec{W} \cup \{X_1\}$, $\vec{w} \cdot \langle x_1' \rangle$, and $\vec{x}$. It is immediate that AC1 holds for $\vec{X}^- = \vec{x}$, and that AC2(a) hold with these witnesses. Thus, AC2(b) must fail. It follows that there must exist some subset $\vec{W}'$ of $\vec{W}$ and subset $\vec{Z}'$ of $\vec{Z}$ such that either (a) $(M, \vec{u}) \models [\vec{X}^- = \vec{x}, X_1 = x_1', \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}^*]\neg\varphi$ or (b) $(M, \vec{u}) \models [\vec{X}^- = \vec{x}, \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}^*]\neg\varphi$. Option (b) cannot hold, because, by (1), it holds iff $(M, \vec{u}) \models [\vec{X} = \vec{x}, \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}^*]\neg\varphi$, which contradicts the assumption that $\vec{X} = \vec{x}$ is a weak cause of φ with $\vec{W}$, $\vec{w}$, and $\vec{x}'$ as witnesses. Thus, (a) must hold. But now it follows that $X_1 = x_1$ is a cause of φ, with $\vec{W} \cup \vec{X}^-$, $\vec{w} \cdot \vec{x}$, and x_1' as witnesses: AC1 and AC3 are immediate, AC2(a) follows from the assumption that (a) holds, and AC2(b) follows from the fact that $\vec{X} = \vec{x}$ is a weak cause with $\vec{W}$, $\vec{w}$, and $\vec{x}'$ as witnesses. ∎

Theorem 5.4: *If $\{X_1 = x_1, \ldots, X_k = x_k\}$ is a cause of φ in M according to the causal NESS test, then $k = 1$.*

Proof: Suppose that $\mathbf{S}$ is a witness of $\{X_1 = x_1, \ldots, X_k = x_k\}$ being a cause of φ in $(M, \vec{u})$ according to the causal NESS test and, by way of contradiction, that $k > 1$. $\mathbf{S}$ is not a witness for $\{X_1 = 1, \ldots, X_{k-1} = x_{k-1}\}$ being a cause of φ (otherwise NT4 would be violated). Thus, it must be the case that $\mathbf{S}' = \mathbf{S} - \{X_1 = 1, \ldots, X_{k-1} = x_{k-1}\}$ is strongly sufficient for φ in $(M, \vec{u})$. But then it follows that that $X_k = x_k$ is a cause of φ in $(M, \vec{u})$ with $\mathbf{S}'$ as a witness. To see this, note that clearly $\mathbf{S}'$ satisfies NT1, since $\mathbf{S}$ does. By assumption, $\mathbf{S}'$ is strongly sufficient for φ in $(M, \vec{u})$, so NT2 holds. And, also by assumption, $\mathbf{S}' - \{X_k = x_k\} = \mathbf{S} - \{X_1 = x_1, \ldots, X_k = x_k\}$ is not a strongly sufficient cause of φ, so NT3 holds. NT4 trivially holds. This shows that $\vec{X} = \vec{x}$ is not a cause of φ according to the causal NESS test, since it does not satisfy NT4. ∎

Theorem 5.5: *Suppose that $X = x$ is a cause of φ in $(M, \vec{u})$ according to the causal NESS test with witness $\mathbf{S}$, and there exists a (possible empty) set T of variables not mentioned in φ or $\mathbf{S}$ and a context $\vec{u}'$ such that the following properties hold:*

SH1. *$\mathbf{S} - \{X = x\}$ is not a sufficient condition for φ in $(M, \vec{u}')$; that is, $(M, \vec{u}') \models [\mathbf{S} - \{X = x\}]\neg\varphi$.*

SH2. *The variables in $\vec{T}$ depend only on the context in $\vec{u}$ and $\vec{u}'$; that is, for all $\vec{t}$, $\vec{T}'$ disjoint from $\vec{T}$, and $\vec{t}'$, we have $(M, \vec{u}) \models \vec{T} = \vec{t}$ iff $(M, \vec{u}) \models [\vec{T}' = \vec{t}'](\vec{T} = \vec{t})$, and similarly for context $\vec{u}'$.*

SH3. *φ is determined by $\vec{T}$ and X in contexts $\vec{u}$ and $\vec{u}'$; that is, for all $\vec{t}$, $\vec{T}'$ disjoint from $\vec{T}$ and X, x', and $\vec{t}'$, we have $(M, \vec{u}') \models [\vec{T} = \vec{t}, \vec{T}' = \vec{t}', X = x']\varphi$ iff $(M, \vec{u}) \models [\vec{T} = \vec{t}, \vec{T}' = \vec{t}', X = x']\varphi$.*

SH4. *In context $\vec{u}$, $\mathbf{S} - \{X = x\}$ depends only on $X = x$ in $\vec{u}$; that is, for all $\vec{T}'$ disjoint from $\mathbf{S}$ and $\vec{t}'$, we have $(M, \vec{u}) \models [\vec{X} = x; \vec{T}' = \vec{t}']\mathbf{S}$.*

Then $X = x$ is a cause of φ in $(M, \vec{u})$ according to the HP definition.

Proof: Suppose that the hypothesis of the proposition holds. By SH1, $(M, \vec{u}') \models [\mathbf{S} - \{X = x\}]\neg\varphi$. Choose x' such that $(M, \vec{u}') \models [\mathbf{S} - \{X = x\}](X = x')$. I claim that we must have $x \neq x'$. For if $(M, \vec{u}') \models [\mathbf{S} - \{X = x\}](X = x)$, then $(M, \vec{u}') \models [\mathbf{S}]\neg\varphi$, contradicting the assumption that $\mathbf{S}$ is strongly sufficient for φ. Let $\vec{W}$ consist of all the variables in $\mathbf{S}$ other than X, together with the set $\vec{T}$ that satisfies SH2 and SH3; let $\vec{Z}$ consist of all the remaining endogenous variables. Let $\vec{w}$ be such that $(M, \vec{u}') \models [\mathbf{S} - \{X = x\}](\vec{W} = \vec{w})$. Note that $\vec{W} = \vec{w}$ subsumes (i.e., includes all the assignments in) $\mathbf{S} - \{X = x\}$. It follows that $(M, \vec{u}') \models [X = x', \vec{W} = \vec{w}]\neg\varphi$. By SH3, we must have $(M, \vec{u}) \models [X = x', \vec{W} = \vec{w}]\neg\varphi$. Thus, AC2(a) holds. For AC2(b), let $\vec{W}'$ be an arbitrary subset of $\vec{W}$ and let $\vec{Z}'$ be an arbitrary subset of $\vec{Z}$. As in the statement of AC2(b), suppose that $(M, \vec{u}) \models \vec{Z} = \vec{z}^*$. We want to show that $(M, \vec{u}) \models [X = x, \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}^*]\varphi$. Let $\vec{T}^* = \vec{T} - \vec{W}'$. Suppose that $(M, \vec{u}) \models \vec{T}^* = \vec{t}^*$. First note that since $\mathbf{S}$ is strongly sufficient for φ in $(M, \vec{u})$, we must have $(M, \vec{u}') \models [\mathbf{S}; \vec{Z}' = \vec{z}^*, \vec{T}^* = \vec{t}^*]\varphi$. Let $\vec{W}'' = \vec{W}' \cap \vec{T}$. Since $\vec{W}'' \subseteq \vec{T}$ and $(M, \vec{u}') \models [\mathbf{S} - \{X = x\}](\vec{W}'' = \vec{w})$, by SH2 we must have $(M, \vec{u}') \models \vec{W}'' = \vec{w}$ and $(M, \vec{u}') \models [\mathbf{S}, \vec{Z}' = \vec{z}^*, \vec{T}^* = \vec{t}^*](\vec{W}'' = \vec{w})$. Thus, $(M, \vec{u}') \models [\mathbf{S}, \vec{W}'' = \vec{w}, \vec{Z}' = \vec{z}^*, \vec{T}^* = \vec{t}^*]\varphi$. Note that all the variables in $\vec{W}' - \vec{W}''$ are in $\mathbf{S} - \{X = x\}$, and they are assigned the same values in $\vec{W}' = \vec{w}$ as in $\mathbf{S}$. Thus, it follows that $(M, \vec{u}') \models [\mathbf{S}, \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}^*, \vec{T}^* = \vec{t}^*]\varphi$. By SH3, $(M, \vec{u}) \models [\mathbf{S}, \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}^*, \vec{T}^* = \vec{t}^*]\varphi$. By SH2, it follows that $(M, \vec{u}) \models [\mathbf{S}, \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}^*]\varphi$. Finally, by SH4, it follows that $(M, \vec{u}) \models [X = x, \vec{Z}' = \vec{z}^*, \vec{W} = \vec{w}]\varphi$, as desired. ∎

Theorem 5.6: *Suppose that $X = x$ is a cause of φ in $(M, \vec{u})$ according to the HP definition, with $\vec{W}$, $\vec{w}$, and x' as witnesses. Suppose that there exists a subset $\vec{W}' \subseteq \vec{W}$ such that $(M, \vec{u}') \models \vec{W}' = \vec{w}$ (that is, the assignment $\vec{W}' = \vec{w}$ does not change the values of the variables in $\vec{W}'$ in context $(M, \vec{u})$) and a context $\vec{u}'$ such that the following conditions hold, where $\vec{W}'' = \vec{W} - \vec{W}'$:*

SN1. $(M, \vec{u}') \models [\vec{W}' = \vec{w}](X = x' \wedge \vec{W}'' = \vec{w})$.

SN2. *$\vec{W}''$ is independent of $\vec{Z}$ given $X = x$ and $\vec{W} = \vec{w}$ in $\vec{u}'$, so that if $\vec{Z}' \subseteq \vec{Z}$, then for all $\vec{z}'$, we have $(M, \vec{u}') \models [X = x, \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}'](\vec{W}'' = \vec{w})$.*

SN3. *φ is independent of $\vec{u}$ and $\vec{u}'$ conditional on X and $\vec{W} = \vec{w}$; that is if $\vec{Z}' \subseteq \vec{Z}$, then for all $\vec{z}'$ and x'', we have $(M, \vec{u}') \models [X = x'', \vec{W} = \vec{w}', \vec{Z} = \vec{z}']\varphi$ iff $(M, \vec{u}) \models [X = x'', \vec{W} = \vec{w}', \vec{Z} = \vec{z}']\varphi$.*

Then $X = x$ is a cause of φ in $(M, \vec{u})$ with respect to $\{\vec{u}, \vec{u}'\}$ according to the causal NESS test.

Proof: Let $\mathbf{S} = \{X = x, \vec{W}' = \vec{w}\}$. Clearly NT1 holds. By assumption, $(M, \vec{u}) \models [X = x', \vec{W} = \vec{w}]\neg\varphi$. By SN3, $(M, \vec{u}') \models [X = x', \vec{W} = \vec{w}]\neg\varphi$. By SN1, it follows that $(M, \vec{u}') \models [\vec{W}' = \vec{w}]\neg\varphi$, so NT3 holds. For NT2, we must show that for all $\vec{Z}' \subseteq \vec{Z} \cup \vec{W}''$, $(M, \vec{u}) \models [X = x, \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}^*]\varphi$, and similarly for $\vec{u}'$. For $\vec{u}$, this is immediate from AC2(b). To see that it also holds for $\vec{u}'$, first note that by AC2(b), we also have $(M, \vec{u}) \models [X = x, \vec{W} = \vec{w}, \vec{Z}'' = \vec{z}^*] \models \varphi$, where $\vec{Z}'' = \vec{Z}' \cap \vec{Z}$. By SN3, $(M, \vec{u}') \models [X = x, \vec{W} = \vec{w}, \vec{Z}'' = \vec{z}^*]\varphi$. By SN2, it follows that $(M, \vec{u}') \models [X = x, \vec{W}' = \vec{w}, \vec{Z}' = \vec{z}^*]\varphi$. Thus, NT2 holds with respect to $\{\vec{u}, \vec{u}'\}$. Clearly NT4 holds, so $X = x$ is a cause of φ in $(M, \vec{u})$ with respect to $\{\vec{u}, \vec{u}'\}$ according to the causal NESS test. ∎