Chapter 2

The HP Definition of Causality

The fluttering of a butterfly’s wing in Rio de Janeiro, amplified by atmospheric currents, could cause a tornado in Texas two weeks later.

Edward Lorenz

There is only one constant, one universal. It is the only real truth. Causality. Action, reaction. Cause and effect.

Merovingian, The Matrix Reloaded

In this chapter, I go through the HP definition in detail. The HP definition is a formal, mathematical definition. Although this does add some initial overhead, it has an important advantage: it prevents ambiguity about whether \( A \) counts as a cause of \( B \). There is no need, as in many other definitions, to try to understand how to interpret the words. For example, recall the INUS condition from the notes in Chapter 1. For \( A \) to be a cause of \( B \) under this definition, \( A \) has to be a necessary part of a condition that is itself unnecessary but insufficient for \( B \). But what is a “condition”? The formalization of INUS suggests that it is a formula or set of formulas. Is there any constraint on this set? What language is it expressed in?

This lack of ambiguity is obviously critical if we want to apply causal reasoning in the law. But as we shall see in Chapter 8, it is equally important in other applications of causality, such as program verification, auditing, and database queries. However, even if there is no ambiguity about the definition, it does not follow that there can be no disagreement about whether \( A \) is a cause of \( B \).

To understand how this can be the case, it is best to outline the general approach. The first step in the HP definition involves building a formal model in which causality can be determined unambiguously. Among other things, the model determines the language that is used to describe the world. We then define only what it means for \( A \) to be a cause of \( B \) in model \( M \). It is possible to construct two closely related models \( M_1 \) and \( M_2 \) such that \( A \) is a cause of \( B \) in \( M_1 \) but not in \( M_2 \). I do not believe that there is, in general, a “right” model; in any case, the definition is silent on what makes one model better than another. (This is an important issue, however. I do think that there are criteria that can help judge whether one
model is better than another; see below and Chapter 4 for more on this point.) Here we already see one instance where, even if there is agreement regarding the definition of causality, we can get disagreements regarding causality: there may be disagreement about which model better describes the real world. This is arguably a feature of the definition. 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.

2.1 Causal Models

The model assumes that the world is described in terms of variables; these variables can take on various 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 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;
- $MD$ for match dropped (by arsonist), where $MD = 1$ if the arsonist dropped a lit match and $MD = 0$ otherwise.

If we are considering a voting scenario where there are eleven voters voting for either Billy or Suzy, we can describe the world using twelve variables, $V_1, \ldots, V_{11}, W$, where $V_i = 0$ if voter $i$ voted for Billy and $V_i = 1$ if voter $i$ voted for Suzy, for $i = 1, \ldots, 11$, $W = 0$ if Billy wins, and $W = 1$ if Suzy wins.

In these two examples, all the variables are binary, that is, they take on only two values. There is no problem allowing a variable to have more than two possible values. For example, the variable $V_i$ could be either 0, 1, or 2, where $V_i = 2$ if $i$ does not vote; similarly, we could take $W = 2$ if the vote is tied, so neither Billy nor Suzy wins.

The choice of 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, if there is no variable corresponding to smoking in model $M$, then in $M$, smoking cannot be a cause of Sam’s lung cancer. Thus, if we want to consider smoking as a potential cause of lung cancer, $M$ is an inappropriate model. (As an aside, the reader may note that here and elsewhere I put “right” in quotes; that is because it is not clear to me that the notion of being “right” is even well defined.)

Some variables may have a causal influence on others. This influence is modeled by a set of structural equations. For example, if we want to model the fact that if the arsonist drops a match or lightning strikes then a fire starts, we could use the variables $MD$, $FF$, and $L$ as above, with the equation $FF = \max(L, MD)$; that is, the value of the variable $FF$ is the maximum of the values of the variables $MD$ and $L$. This equation says, among other things, that if $MD = 0$ and $L = 1$, then $FF = 1$. The equality sign in this equation should be thought of more like an assignment statement in programming languages; once we set the values of $MD$ and $L$, the value of $FF$ is set to their maximum. However, despite the equality, a forest fire starting some other way does not force the value of either $MD$ or $L$ to be 1.
2.1 Causal Models

Alternatively, if we want to model the fact that a fire requires both a lightning strike and a dropped match (perhaps the wood is so wet that it needs two sources of fire to get going), then the only change in the model is that the equation for \( FF \) becomes \( FF = \min(L, MD) \); the value of \( FF \) is the minimum of the values of \( MD \) and \( L \). The only way that \( FF = 1 \) is if both \( L = 1 \) and \( MD = 1 \).

Just a notational aside before going on: I sometimes identify binary variables with primitive propositions in propositional logic. And, as in propositional logic, the symbols \( \land \), \( \lor \), and \( \neg \) are used to denote conjunction, disjunction, and negation, respectively. With that identification, instead of writing \( \max(L, MD) \), I write \( L \lor MD \); instead of writing \( \min(L, MD) \), I write \( L \land MD \); and instead of writing \( 1 - L \) or \( 1 - MD \), I write \( \neg L \) or \( \neg MD \). Most people seem to find the logic notation easier to absorb. I hope that the intention will be clear from context.

Going on with the forest-fire example, it is clear that both of these models are somewhat simplistic. Lightning does not always result in a fire, nor does dropping a lit match. One way of dealing with this would be to make the assignment statements probabilistic. For example, we could say that the probability that \( FF = 1 \) conditional on \( L = 1 \) is \( .8 \). I discuss this approach in more detail in Section 2.5. It is much simpler to think of all the equations as being deterministic and then use enough variables to capture all the conditions that determine whether there is a forest fire. One way to do this is simply to add those variables explicitly. For example, we could add variables that talk about the dryness of the wood, the amount of undergrowth, the presence of sufficient oxygen (fires do not start so easily on the top of high mountains), and so on. If a modeler does not want to add all these variables explicitly (the details may simply not be relevant to the analysis), then another alternative is to use a single variable, say \( U \), which intuitively incorporates all the relevant factors, without describing them explicitly. The value of \( U \) would determine whether the lightning strikes and whether the match is dropped by the arsonist.

The value of \( U \) could also determine whether both the match and lightning are needed to start a fire or just one of them suffices. For simplicity, rather than using \( U \) in this way, I consider two causal models. In one, called the conjunctive model, both the match and lightning are needed to start the forest fire; in the other, called the disjunctive model, only one is needed. In each of these models, \( U \) determines only whether the lightning strikes and whether the match is dropped. Thus, I assume that \( U \) can take on four possible values of the form \((i,j)\), where \( i \) and \( j \) are each either 0 or 1. Intuitively, \( i \) describes whether the external conditions are such that the lightning strikes (and encapsulates all such conditions, e.g., humidity and temperature), and \( j \) describes whether the arsonist drops the match (and thus encapsulates all the psychological conditions that determine this). For future reference, let \( U_1 \) and \( U_2 \) denote the components of the value of \( U \) in this example, so that if \( U = (i, j) \), then \( U_1 = i \) and \( U_2 = j \).

Here I have assumed a single variable \( U \) that determines the values of both \( L \) and \( MD \). I could have split \( U \) into two variables, say \( U_L \) and \( U_{MD} \), where \( U_L \) determines the value of \( L \) and \( U_{MD} \) determines the value of \( MD \). It turns out that whether we use a single variable with four possible values or two variables, each with two values, makes no real difference. (Other modeling choices can make a big difference—I return to this point below.)

It is reasonable to ask why I have chosen to describe the world in terms of variables and their values, related via structural equations. Using variables and their values is quite standard
in fields such as statistics and econometrics for good reason; it is a natural way to describe situations. The examples later in this section should help make that point. It is also quite close to propositional logic; we can think of a primitive proposition in classical logic as a binary variable, whose values are either true or false. Although there may well be other reasonable ways of representing the world, this seems like a useful approach.

As for the use of structural equations, recall that my goal is to give a definition of actual causality in terms of counterfactuals. Certainly if the but-for test applies, so that, but for $A$, $B$ would not have happened, I want the definition to declare that $A$ is a cause of $B$. That means I need a model rich enough to capture the fact that, but for $A$, $B$ would not have happened. Because I model the world in terms of variables and their values, the $A$ and the $B$ in the definition will be statements such as $X = x$ and $Y = y$. Thus, I want to say that if $X$ had taken on some value $x'$ different from $x$, then $Y$ would not have had value $y$. To do this, I need a model that makes it possible to consider the effect of intervening on $X$ and changing its value from $x$ to $x'$. Describing the world in terms of variables and their values makes it easy to describe such interventions; using structural equations makes it easy to define the effect of an intervention.

This can be seen already in the forest-fire example. In the conjunctive model, if we see a forest fire, then we can say that if the arsonist hadn’t dropped a match, there would have been no forest fire. This follows in the model because setting $MD = 0$ makes $FF = \min(L, MD) = 0$.

Before going into further details, I give an example of the power of structural equations. Suppose that we are interested in the relationship between education, an employee’s skill level, and his salary. Suppose that education can affect skill level (but does not necessarily always do so, since a student may not pay attention or have poor teachers) and skill level affects salary. We can give a simplified model of this situation using four variables:

- $E$ for education level, with values 0 (no education), 1 (one year of education), and 2 (2 years of education);
- $SL$ for skill level, with values 0 (low), 1 (medium), and 2 (high);
- $S$ for salary, which for simplicity we can also take to have three values: 0 (low), 1 (medium), and 2 (high);
- $U$, a variable that determines whether education will have an impact on skill level.

There are two relevant equations. The first says that education determines skill level, provided that the external conditions (determined by $U$) are favorable, and otherwise has no impact:

$$SL = \begin{cases} E & \text{if } U = 1 \\ 0 & \text{if } U = 0. \end{cases}$$

The second says that skill level determines salary:

$$S = SL.$$
level 1. We can thus infer that $U = E = SL = 1$. We can ask what Fred’s salary level would have been had he attained education level 2. This amounts to intervening and setting $E = 2$. Since $U = 1$, it follows that $SL = 2$ and thus $S = 2$; Fred would have a high salary. Similarly, if we observe that Fred’s education level is 1 but his salary is low, then we can infer that $U = 0$ and $SL = 0$, and we can say that his salary would still be low even if he had an extra year of education.

A more sophisticated model would doubtless include extra factors and more sensitive dependencies. Still, even at this level, I hope it is clear that the structural equations framework can give reasonable answers.

When working with structural equations, it turns out to be conceptually useful to split the variables into two sets: the exogenous variables, whose values are determined by factors outside the model, and the endogenous variables, whose values are ultimately determined by the exogenous variables. In the forest-fire example, the variable $U$, which determines whether the lightning strikes and whether the arsonist drops a match, is exogenous; the variables $FF$, $L$, and $MD$ are endogenous. The value of $U$ determines the values of $L$ and $MD$, which in turn determine the value of $FF$. We have structural equations for the three endogenous variables that describe how this is done. In general, there is a structural equation for each endogenous variable, but there are no equations for the exogenous variables; as I said, their values are determined by factors outside the model. That is, the model does not try to “explain” the values of the exogenous variables; they are treated as given.

The split between exogenous and endogenous variables has another advantage. The structural equations are all deterministic. As we shall see in Section 2.5, when we want to talk about the probability that $A$ is a cause of $B$, we can do so quite easily by putting a probability distribution on the values of the exogenous variables. This gives us a way of talking about the probability of there being a fire if lightning strikes, while still using deterministic equations.

In any case, with this background, we can formally define a causal model $M$ as 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 structural equations, relating the values of the variables. In the next two paragraphs, I define $\mathcal{S}$ and $\mathcal{F}$ formally; the definitions can be skipped by the less mathematically inclined reader.

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 $Y$ (i.e., the set of values over which $Y$ ranges). As suggested earlier, in the forest-fire example, we can take $\mathcal{U} = \{U\}$; that is, $U$ is exogenous, $\mathcal{R}(U)$ consists of the four possible values of $U$ discussed earlier, $\mathcal{V} = \{FF, L, MD\}$, and $\mathcal{R}(FF) = \mathcal{R}(L) = \mathcal{R}(MD) = \{0, 1\}$.

The function $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted $F_X$ such that $F_X$ maps $\times_{Z \in (\mathcal{U} \cup \mathcal{V} - \{X\})} \mathcal{R}(Z)$ to $\mathcal{R}(X)$. (Recall that $\mathcal{U} \cup \mathcal{V} - \{X\}$ is the set consisting of all the variables in either $\mathcal{U}$ or $\mathcal{V}$ that are not in $\{X\}$. The notation $\times_{Z \in (\mathcal{U} \cup \mathcal{V} - \{X\})} \mathcal{R}(Z)$ denotes the cross-product of $\mathcal{R}(Z)$, where $Z$ ranges over the variables in $\mathcal{U} \cup \mathcal{V} - \{X\}$; thus, if $\mathcal{U} \cup \mathcal{V} - \{X\} = \{Z_1, \ldots, Z_k\}$, then $\times_{Z \in (\mathcal{U} \cup \mathcal{V} - \{X\})} \mathcal{R}(Z)$ consists of tuples of the form $(z_1, \ldots, z_k)$, where $z_i$ is a possible value of $Z_i$, for $i = 1, \ldots, k$.) 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 $\mathcal{U} \cup \mathcal{V}$. In the running forest-fire example, $F_{FF}$ would depend on whether
Chapter 2. The HP Definition of Causality

we are considering the conjunctive model (where both the lightning strike and dropped match are needed for the fire) or the disjunctive model (where only one of the two is needed). In the conjunctive model, as I have already noted, \( F_{FF}(L, MD, U) = 1 \) iff ("iff" means "if and only if"—this standard abbreviation is used throughout the book) \( \min(L, MD) = 1 \); in the disjunctive model, \( F_{FF}(L, MD, U) = 1 \) iff \( \max(L, MD) = 1 \). Note that, in both cases, the value of \( F_{FF} \) is independent of \( U \); it depends only on the values of \( L \) and \( MD \). Put another way, this means that when we hold the values of \( L \) and \( MD \) fixed, changing the value of \( U \) has no effect on the value of \( F_{FF} \). While the values of \( MD \) and \( L \) do depend on the value of \( U \), writing things in this way allows us to consider the effects of external interventions that may override \( U \). For example, we can consider \( F_{FF}(0, 1, (1, 1)) \); this tells us what happens if the lightning does not strike and the match is dropped, even if the value of \( U \) is such that \( L = 1 \). In what follows, we will be interested only in interventions on endogenous variables. As I said, we take the values of the exogenous variables as given, so do not intervene on them.

The notation \( F_X(Y, Y', U) \) for the equation describing how the value of \( X \) depends on \( Y \), \( Y' \), and \( U \) is not terribly user-friendly. I typically simplify notation and write \( X = Y + U \) instead of \( F_X(Y, Y', U) = Y + U \). (Note that the variable \( Y' \) does not appear on the right-hand side of the equation. That means that the value of \( X \) does not depend on that of \( Y' \).) With this simplified notation, the equations for the forest-fire example are \( L = U_1 \), \( MD = U_2 \), and either \( FF = \min(L, MD) \) or \( FF = \max(L, MD) \), depending on whether we are considering the conjunctive or disjunctive model.

Although I may write something like \( X = Y + U \), the fact that \( X \) is assigned \( Y + U \) does not imply that \( Y \) is assigned \( X - U \); that is, \( F_Y(X, Y', U) = X - U \) does not necessarily hold. The equation \( X = Y + U \) implies that if \( Y = 3 \) and \( U = 2 \), then \( X = 5 \), regardless of how \( Y' \) is set. Going back to the forest-fire example, setting \( FF = 0 \) does not mean that the match is "undropped"! This asymmetry in the treatment of the variables on the left- and right-hand sides of the equality sign means that (the value of) \( Y \) can depend on (the value of) \( X \) without \( X \) depending on \( Y \). This, in turn, is why causality is typically asymmetric in the formal definition that I give shortly: if \( A \) is a cause of \( B \), then \( B \) will typically not be a cause of \( A \). (See the end of Section 2.7 for more discussion of this issue.)

As I suggested earlier, the key role of the structural equations is that they allow us to determine what happens if things had been other than they were, perhaps due to an external intervention; for example, the equations tell us what would happen if the arsonist had not dropped a match (even if in fact he did). This will be critical in defining causality.

Since a world in a causal model is described by the values of variables, understanding what would happen if things were other than they were amounts to asking what would happen if some variables were set to values perhaps different from their actual values. Setting the value of some variable \( X \) to \( x \) in a causal model \( M = (S, F) \) results in a new causal model denoted \( M_{X\leftarrow x} \). In the new causal model, the equation for \( X \) is simple: \( X \) is just set to \( x \); the remaining equations are unchanged. More formally, \( M_{X\leftarrow x} = (S, F_{X\leftarrow x}) \), where \( F_{X\leftarrow x} \) is the result of replacing the equation for \( X \) in \( F \) by \( X = x \) and leaving the remaining equations untouched. Thus, if \( M^C \) is the conjunctive model of the forest fire, then in \( M^C_{MD\leftarrow 0} \), the model that results from intervening in \( M^C \) by having the arsonist not drop a match, the equation \( MD = U_2 \) is replaced by \( MD = 0 \).
As I said in Chapter 1, the HP definition involves counterfactuals. The equations in a causal model can be given a straightforward counterfactual interpretation. An equation such as \( x = F_X(u, y) \) should be thought of as saying that in a context where the exogenous variable \( U \) has value \( u \), if \( Y \) were set to \( y \) by some means (not specified in the model), then \( X \) would take on the value \( x \). As I noted earlier, when \( Y \) is set to \( y \), this can override the value of \( Y \) according to the equations. For example, \( MD \) can be set to 0 even in a context where the exogenous variable \( U = (1, 1) \), so \( MD \) would be 1.

It may seem somewhat circular to use causal models, which clearly already encode causal relationships, to define causality. There is some validity to this concern. After all, how do we determine whether a particular equation holds? We might believe that dropping a lit match results in a forest burning in part because of our experience with lit matches and dry wood, and thus believe that a causal relationship holds between the lit match and the dry wood. We might also have a general theory of which this is a particular outcome, but there too, roughly speaking, the theory is being understood causally. Nevertheless, I would claim that this definition is useful. In many examples, there is general agreement as to the appropriate causal model. The structural equations do not express actual causality; rather, they express the effects of interventions or, more generally, of variables taking on values other than their actual values.

Of course, there may be uncertainty about the effects of interventions, just as there may be uncertainty about the true setting of the values of the exogenous variables in a causal model. For example, we may be uncertain about whether smoking causes cancer (this represents uncertainty about the causal model), uncertain about whether a particular patient Sam actually smoked (this is uncertainty about the value of the exogenous variables that determine whether Sam smokes), and uncertain about whether Sam’s brother Jack, who did smoke and got cancer, would have gotten cancer had he not smoked (this is uncertainty about the effect of an intervention, which amounts to uncertainty about the values of exogenous variables and possibly the equations). All this uncertainty can be described by putting a probability on causal models and on the values of the exogenous variables. We can then talk about the probability that \( A \) is a cause of \( B \). (See Section 2.5 for further discussion of this point.)

Using (the equations in) a causal model, we can determine whether a variable \( Y \) is (in)dependent of variable \( X \). \( Y \) depends on \( X \) if there is some setting of all the variables in \( U \cup V \) other than \( X \) and \( Y \) such that varying the value of \( X \) in that setting results in a variation in the value of \( Y \); that is, there is a setting \( \vec{z} \) of the variables other than \( X \) and \( Y \) and values \( x \) and \( x' \) of \( X \) such that \( F_Y(x, \vec{z}) \neq F_Y(x', \vec{z}) \). (The \( \vec{z} \) represents the values of all the other variables. This vector notation is a useful shorthand that is used throughout the book; I say more about it for readers unfamiliar with it at the end of the section.) Note that by “dependency” here I mean something closer to “immediate dependency” or “direct dependency”. This notion of dependency is not transitive; that is, if \( X_1 \) depends on \( X_2 \) and \( X_2 \) depends on \( X_3 \), then it is not necessarily the case that \( X_1 \) depends on \( X_3 \). If \( Y \) does not depend on \( X \), then \( Y \) is independent of \( X \).

Whether \( Y \) depends on \( X \) may depend in part on the variables in the language. In the original description of the voting scenario, we had just twelve variables, and \( W \) depended on each of \( V_1, \ldots, V_{11} \). However, suppose that the votes are tabulated by a machine. We can add a variable \( T \) that describes that final tabulation of the machine, where \( T \) can have any value
of the form \((t_1, t_2)\), where \(t_1\) represents the number of votes for Billy and \(t_2\) represents the number of votes for Suzy (so that \(t_1\) and \(t_2\) are non-negative integers whose sum is 11). Now \(W\) depends only on \(T\), while \(T\) depends on \(V_1, \ldots, V_{11}\).

In general, every variable can depend on every other variable. But in most interesting situations, each variable depends on relatively few other variables. The dependencies between variables in a causal model \(M\) can be described using a causal network (sometimes called a causal graph), consisting of nodes and directed edges. I often omit the exogenous variables from the graph. Thus, both Figure 2.1(a) (where the exogenous variable is included) and Figure 2.1(b) (where it is omitted) would be used to describe the forest-fire example. Figure 2.1(c) is yet another description of the forest-fire example, this time replacing the single exogenous variable \(U\) by an exogenous variable \(U_1\) that determines \(L\) and an exogenous variable \(U_2\) that determines \(MD\). Again, Figure 2.1(b) is the causal graph that results when the exogenous variables in Figure 2.1(c) are omitted. Since all the “action” happens with the endogenous variables, Figure 2.1(b) really gives us all the information we need to analyze this example.

The fact that there is a directed edge from \(U\) to both \(L\) and \(MD\) in Figure 2.1(a) (with the direction marked by the arrow) says that the value of the exogenous variable \(U\) affects the value of \(L\) and \(MD\), but nothing else affects it. The directed edges from \(L\) and \(MD\) to \(FF\) say that only the values of \(MD\) and \(L\) directly affect the value of \(FF\). (The value of \(U\) also affects the value of \(FF\), but it does so indirectly, through its effect on \(L\) and \(MD\).)

Causal networks convey only the qualitative pattern of dependence; they do not tell us how a variable depends on others. For example, the same causal network would be used for both the conjunctive and disjunctive models of the forest-fire example. Nevertheless, causal networks are useful representations of causal models.

I will be particularly interested in causal models where there are no circular dependencies. For example, it is not the case that \(X\) depends on \(Y\) and \(Y\) depends \(X\) or, more generally, that \(X_1\) depends on \(X_2\), which depends on \(X_3\), which depends on \(X_4\), which depends on \(X_1\). Informally, a model is said to be recursive (or acyclic) if there are no such dependency cycles (sometimes called feedback cycles) among the variables.
2.1 Causal Models

The intuition that the values of exogenous variables are determined by factors outside the model is formalized by requiring that an exogenous variable does not depend on any other variables. (Actually, nothing in the following discussion changes if we allow an exogenous variable to depend on other exogenous variables.) In a recursive model, we can say more about how the values of endogenous are determined. Specifically, the values of some of the endogenous variables depend only on the values of the exogenous variables. This is the case about how the values of endogenous are determined. Specifically, the values of some of the variables does not depend on any endogenous variables.) In a recursive model, we can say more about the context and the values of the first-level endogenous variables. The values of the “second-level” endogenous variables depend only on the values of exogenous variables and the values of first-level endogenous variables. The variable \( FF \) is a second-level variable in this sense; its value is determined by that of \( L \) and \( MD \), and these are first-level variables. We can then define third-level variables, fourth-level variables, and so on.

Actually, the intuition I have just given is for what I call a strongly recursive model. It is easy to see that in a strongly recursive model, the values of all variables are determined given a context, that is, a setting \( \vec{u} \) for the exogenous variables in \( \mathcal{U} \). Given \( \vec{u} \), we can determine the values of the first-level variables using the equations; we can then determine the values of the second-level variables (whose values depend only on the context and the values of the first-level variables), then the third-level variables, and so on. The notion of a recursive model generalizes the definition of a strongly recursive model (every strongly recursive model is recursive, but the converse is not necessarily true) while still keeping this key feature of having the context determine the values of all the endogenous variables. In a recursive model, the partition of endogenous variables into first-level variables, second-level variables, and so on can depend on the context; in different contexts, the partition might be different. Moreover, in a recursive model, our definition guarantees that causality is asymmetric: it cannot be the case that \( A \) is a cause of \( B \) and \( B \) is a cause of \( A \) if \( A \) and \( B \) are distinct. (See the discussion at the end of Section 2.7.)

The next four paragraphs provide a formal definition of recursive models. They can be skipped on a first reading.

Say that a model \( M \) is strongly recursive (or acyclic) if there is some partial order \( \preceq \) on the endogenous variables in \( M \) (the ones in \( \mathcal{V} \)) such that unless \( X \preceq Y \), \( Y \) is not affected by \( X \). Roughly speaking, \( X \preceq Y \) denotes that \( X \) affects \( Y \) (think of “affects” here as the transitive closure of the direct dependency relation discussed earlier; \( X \) affects \( Y \) if there is some chain \( X_1, \ldots, X_k \) such that \( X = X_1, Y = X_k \), and \( X_{i+1} \) directly depends on \( X_i \) for \( i = 1, \ldots, k - 1 \). The fact that \( \preceq \) is a partial order means that \( \preceq \) is a reflexive, anti-symmetric, and transitive relation. Reflexivity means that \( X \preceq X \) for each variable \( X \)—\( X \) affects \( X \); anti-symmetry means that if \( X \preceq Y \) and \( Y \preceq X \), then \( X = Y \)—if \( X \) affects \( Y \) and \( Y \) affects \( X \), then we must have \( X = Y \); finally, transitivity means that if \( X \preceq Y \) and \( Y \preceq Z \), then \( X \preceq Z \)—if \( X \) affects \( Y \) and \( Y \) affects \( Z \), then \( X \) affects \( Z \). Since \( \preceq \) is partial, it may be the case that for some variables \( X \) and \( Y \), neither \( X \preceq Y \) nor \( Y \preceq X \) holds; that is, \( X \) does not affect \( Y \) and \( Y \) does not affect \( X \).

While reflexivity and transitivity seem to be natural properties of the “affects” relation, anti-symmetry is a nontrivial assumption. The fact that \( \preceq \) is anti-symmetric and transitive means that there is no cycle of dependence between a collection \( X_1, \ldots, X_n \) of variables. It
cannot be the case that $X_1$ affects $X_2$, $X_2$ affects $X_3$, ..., $X_{n-1}$ affects $X_n$, and $X_n$ affects $X_1$. For then we would have $X_1 \preceq X_2$, $X_2 \preceq X_3$, ..., $X_{n-1} \preceq X_n$, and $X_n \preceq X_1$. By transitivity, we would have $X_2 \preceq X_1$, violating anti-symmetry. A causal network corresponding to a causal model where there is no such cyclic dependence between the variables is acyclic. That is, there is no sequence of directed edges that both starts and ends at the same node.

A model $M$ is recursive if, for each context (setting of the exogenous variables) $\vec{u}$, there is a partial order $\preceq_{\vec{u}}$ of the endogenous variables such that unless $X \preceq_{\vec{u}} Y$, $Y$ is independent of $X$ in $(M, \vec{u})$, where $Y$ is independent of $X$ in $(M, \vec{u})$ if, for all settings $\vec{z}$ of the endogenous variables other than $X$ and $Y$, and all values $x$ and $x'$ of $X$, $F_Y(x, \vec{z}, \vec{u}) = F_Y(x', \vec{z}, \vec{u})$. If $M$ is a strongly recursive model, then we can assume that all the partial orders $\preceq_{\vec{u}}$ are the same; in a recursive model, they may differ. Example 2.3.3 below shows why it is useful to consider the more general notion of recursive model.

As I said, if $M$ is a recursive causal model, then given a context $\vec{u}$, there is a unique solution for all the equations. We simply solve for the variables in the order given by $\prec_{\vec{u}}$ (where $X \prec_{\vec{u}} Y$ if $X \preceq_{\vec{u}} Y$ and $X \neq Y$). The value of the variables that come first in the order, that is, the variables $X$ such that there is no variable $Y$ such that $Y \prec_{\vec{u}} X$, depend only on the exogenous variables, so their value is immediately determined by the values of the exogenous variables. The values of variables later in the order can be determined from the equations once we have determined the values of all the variables earlier in the order.

With the definition of recursive model under our belt, we can get back to the issue of choosing the “right” model. There are many nontrivial decisions to be made when choosing the causal 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 forest-fire 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 as given. 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 as given 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). (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.) By making $D$ exogenous, its value is assumed to be determined by external factors that are not being modeled. We could also take the amount of oxygen to be described by an exogenous variable (e.g., 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 variable $MD$ (match dropped by arsonist) and another variable $WB$ (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 $MD$. 


By setting $F_{WB}(1, 1) = 1$, we are saying that the wood will burn if the lit match is dropped and the wood is dry. Thus, the equation is implicitly modeling our assumption that there is sufficient oxygen for the wood to burn. Whether the modeler should include $O$ in the model depends on whether the modeler wants to consider contexts where the value of $O$ is 0. If in all contexts relevant to the modeler there is sufficient oxygen, there is no point in cluttering up the model by adding the variable $O$ (although adding it will not affect anything).

According to the definition of causality in Section 2.2, 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 in that model. 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. Interestingly, in the law, there is a distinction between what are called conditions and causes. Under typical circumstances, the presence of oxygen would be considered a condition, and would thus not count as a cause of the forest burning, whereas the lightning would. While the of oxygen would be considered a condition and would thus not count as a cause of the forest burning, while the lightning would. Although the distinction is considered important, it does not seem to have been carefully formalized. One way of understanding it is in terms of exogenous versus endogenous variables: conditions are exogenous, (potential) causes are endogenous. I discuss an alternative approach to understanding the distinction, in terms of theories of normality, in Section 3.2.

It is not always straightforward to decide what the “right” causal model is in a given situation. What is the “right” set of variables to use? Which should be endogenous and which should be exogenous? As we shall see, different choices of endogenous variables can lead to different conclusions, although they seem to be describing exactly the same situation. And even after we have chosen the variables, how do we determine the equations that relate them? Given two causal models, how do we decide which is better?

These are all important questions. For now, however, I will ignore them. The definition of causality I give in the next section is relative to a model $M$ and context $\vec{u}$. $A$ may be a cause of $B$ relative to $(M, \vec{u})$ and not a cause of $B$ relative to $(M', \vec{u}')$. Thus, the choice of model can have a significant impact in determining causality ascriptions. The definition is agnostic regarding the choice of model; it has nothing to say about whether the choice of variables or the structural equations used are in some sense “reasonable”. Of course, people may legitimately disagree about how well a particular causal model describes the world. Although 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. That said, I do return to some of the issues raised earlier when discussing the examples in this section, and I discuss them in more detail in Chapter 4.

As promised, I conclude this section with a little more discussion of the vector notation, for readers unfamiliar with it. I use this notation throughout the book to denote sets of variables or their values. For example, if there are three exogenous variables, say $U_1$, $U_2$, and $U_3$, and $U_1 = 0$, $U_2 = 0$, and $U_3 = 1$, then I write $\vec{U} = (0, 0, 1)$ as an abbreviation of $U_1 = 0$, $U_2 = 0$, and $U_3 = 1$. This vector notation is also used to describe the values $\vec{x}$ of a collection
Chapter 2. The HP Definition of Causality

\( \vec{X} \) of endogenous variables. I deliberately abuse notation at times and view \( \vec{U} \) both as a set of variables and as a sequence of variables. For example, when I write \( \vec{U} = (0, 0, 1) \), then I am thinking of \( \vec{U} \) as the sequence \((U_1, U_2, U_3)\), and this equality means that \( U_1 = 0, U_2 = 0, \) and \( U_3 = 1 \). However, in expressions such as \( U_2 \in \vec{U} \) and \( \vec{U} \subseteq \vec{U}' \), \( \vec{U} \) should be understood as the corresponding unordered set \( \{U_1, U_2, U_3\} \). I hope that the intent is always clear from context. For the most part, readers who do not want to delve into details can just ignore the vector arrows.

2.2 A Formal Definition of Actual Cause

In this section, I give the HP definition of causality and show how it works in a number of examples. But before giving the definition, I have to define a formal language for describing causes.

2.2.1 A language for describing causality

To make the definition of actual causality precise, it is helpful to have a formal language for making statements about causality. The language is an extension of propositional logic, where the primitive events (i.e., primitive propositions) have the form \( X = x \) for an endogenous variable \( X \) and some possible value \( x \) of \( X \). The primitive event \( MD = 0 \) says “the lit match is not dropped”; similarly, the primitive event \( L = 1 \) says “lightning occurred”. These primitive events can be combined using standard propositional connectives, such as \( \land, \lor, \) and \( \neg \). Thus, the formula \( MD = 0 \lor L = 1 \) says “either the lit match is not dropped or lightning occurred”, \( MD = 0 \land L = 1 \) says “the lit match is not dropped and lightning occurred”, and \( \neg(L = 1) \) says “lightning did not occur” (which is equivalent to \( L = 0 \), given that the only possible values of \( L \) are 0 or 1). A Boolean combination of primitive events is a formula obtained by combining primitive events using \( \land, \lor, \) and \( \neg \). For example, \( \neg(MD = 0 \lor L = 1) \land WB = 1 \) is a Boolean combination of the primitive events \( MD = 0, L = 1, \) and \( WB = 1 \). The key novel feature in the language is that we can talk about interventions. A formula of the form 
\[
[\vec{Y} \leftarrow \vec{y}](X = x)
\]

says that after intervening to set the variables in \( \vec{Y} \) to \( \vec{y} \), \( X \) takes on the value \( x \).

(A short aside: I am thinking of an event here as a subset of a state space; this is the standard usage in computer science and probability. Once I give semantics to causal formulas, it will be clear that we can associate a causal formula \( \varphi \) with a set of contexts—the set of context where \( \varphi \) is true—so it can indeed be identified with an event in this sense. In the philosophy literature, the use of the word “event” is somewhat different, and what counts as an event is contentious; see the notes at the end of this chapter and at the end of Chapter 4.)

Now for the formal details. Given a signature \( S = (U, V, R) \), a primitive event is a formula of the form \( X = x \), for \( X \in V \) and \( x \in R(X) \). A causal formula (over \( S \)) is one of the form 
\[
[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k] \varphi,
\]

where

- \( \varphi \) is a Boolean combination of primitive events,
- \( Y_1, \ldots, Y_k \) are distinct variables in \( V \), and
2.2 A Formal Definition of Actual Cause

- \( y_i \in \mathcal{R}(Y_i) \).

Such a formula is abbreviated as \( \hat{Y} \leftarrow \hat{y} \varphi \), using the vector notation. The special case where \( k = 0 \) is abbreviated as \( [] \varphi \) or, more often, just \( \varphi \). Intuitively, \( [Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k] \varphi \) says that \( \varphi \) would hold if \( Y_i \) were set to \( y_i \), for \( i = 1, \ldots, k \). It is worth emphasizing that the commas in \( [Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k] \varphi \) are really acting like conjunctions; this says that if \( Y_1 \) were set to \( y_1 \) and \( \ldots \) and \( Y_k \) were set to \( y_k \), then \( \varphi \) would hold. I follow convention and use the comma rather than \( \land \). I write \( Y_j \leftarrow y_j \) rather than \( Y_j = y_j \) to emphasize that here \( Y_j \) is assigned the value \( y_j \), as the result of an intervention. For \( S = (\mathcal{U}, \mathcal{V}, \mathcal{R}) \), let \( \mathcal{L}(S) \) consist of all Boolean combinations of causal formulas, where the variables in the formulas are taken from \( \mathcal{V} \) and the sets of possible values of these variables are determined by \( \mathcal{R} \).

The language above lets us talk about interventions. What we need next is a way of defining the semantics of causal formulas: that is, a way of determining when a causal formula is true. A causal formula \( \psi \) is true or false in a causal model, given a context. I call a pair \((M, \bar{u})\) consisting of a causal model \( M \) and context \( \bar{u} \) a causal setting. I write \( (M, \bar{u}) \models \psi \) if the causal formula \( \psi \) is true in the causal setting \((M, \bar{u})\). For now, I restrict attention to recursive models, where, given a context, there are no cyclic dependencies. In a recursive model, \((M, \bar{u}) \models X = x \) if the value of \( X \) is \( x \) once we set the exogenous variables to \( \bar{u} \).

Here I am using the fact that, in a recursive model, the values of all the endogenous variables are determined by the context. If \( \psi \) is an arbitrary Boolean combination of primitive events, then whether \( (M, \bar{u}) \models \psi \) can be determined by the usual rules of propositional logic. For example, \((M, \bar{u}) \models X = x \lor Y = y \) if either \((M, \bar{u}) \models X = x \) or \((M, \bar{u}) \models Y = y \).

Using causal models makes it easy to give the semantics of formulas of the form \( [\hat{Y} \leftarrow \hat{y}](X = x) \) and, more generally, \( [\hat{Y} \leftarrow \hat{y}] \psi \). Recall that the latter formula says that, after intervening to set the variables in \( \hat{Y} \) to \( \hat{y} \), \( \psi \) holds. Given a model \( M \), the model that describes the result of this intervention is \( M_{\hat{Y} \leftarrow \hat{y}} \). Thus,

\[
(M, \bar{u}) \models [\hat{Y} \leftarrow \hat{y}] \psi \iff (M_{\hat{Y} \leftarrow \hat{y}}, \bar{u}) \models \psi.
\]

The mathematics just formalizes the intuition that the formula \( [\hat{Y} \leftarrow \hat{y}] \psi \) is true in the causal setting \((M, \bar{u})\) exactly if the formula \( \psi \) is true in the model that results from the intervention, in the same context \( \bar{u} \).

For example, if \( M^d \) is the disjunctive model for the forest fire described earlier, then \((M^d, (1,1)) \models [MD \leftarrow 0](FF = 1)\): even if the arsonist is somehow prevented from dropping the match, the forest burns (thanks to the lightning); that is, \((M^d_{MD \leftarrow 0}, (1,1)) \models FF = 1 \). Similarly, \((M^d, (1,1)) \models [L \leftarrow 0](FF = 1)\). However, \((M^d, (1,1)) \models [L \leftarrow 0; MD \leftarrow 0](FF = 0)\): if the arsonist does not drop the lit match and the lightning does not strike, then the forest does not burn.

The notation \((M, \bar{u}) \models \varphi \) is standard in the logic and philosophy communities. It is, unfortunately, not at all standard in other communities, such as statistics and econometrics. Although notation is not completely standard in these communities, the general approach has been to suppress the model \( M \) and make the context \( \bar{u} \) an argument of the endogenous variables. Thus, for example, instead of writing \((M, \bar{u}) \models X = x \), in these communities they would write \( X(\bar{u}) = x \). (Sometimes the exogenous variables are also suppressed, or taken as given, so just \( X = x \) is written.) More interestingly, instead of
Chapter 2. The HP Definition of Causality

(M, \vec{u}) \models [X \leftarrow x](Y = y), in these communities they would write \( Y_x(\vec{u}) = y \). Although the latter notation, which I henceforth call statisticians’ notation (although it is more widespread), is clearly more compact, the compactness occasionally comes at the price of clarity. For example, in the disjunctive forest-fire example, what does \( FF(1, 1) = 1 \) mean? Is it \( (Md, (1, 1)) \models [MD \leftarrow 0](FF = 1) \) or \( (Md, (1, 1)) \models [L \leftarrow 0](FF = 1) \)? This problem can be solved by adding the variable being intervened on to the subscript if it is necessary to remove ambiguity, writing, for example, \( FF_{L \leftarrow 0}(1, 1) = 0 \). Things get more complicated when we want to write \( (M, \vec{u}) \models [X \leftarrow x]\phi \), where \( \phi \) is a Boolean combination of primitive events, or when there are several causal models in the picture and it is important to keep track of them. That said, there are many situations when writing something like \( Y_x(\vec{u}) = y \) is completely unambiguous, and it is certainly more compact. To make things easier for those used to this notation, I translate various statements into statisticians’ notation, particularly those where it leads to more compact formulations. This will hopefully have the added advantage of making both notations familiar to all communities.

2.2.2 The HP definition of actual causality

I now give the HP definition of actual causality.

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

Note that this means that the relata of causality (what can be a cause and what can be caused) depend on the language. It is also worth remarking that there is a great deal of debate in the philosophical literature about the relata of causality. Although it is standard in that literature to take the relata of causality to be events, exactly what counts as an event is also a matter of great debate. I do not delve further into these issues here; see the notes at the end of the chapter for references and a little more discussion.

Although I have been talking up to now about “the” HP definition of causality, I will actually consider three definitions, all with the same basic structure. Judea Pearl and I started with a relatively simple definition that gradually became more complicated as we discovered examples that our various attempts could not handle. One of the examples was discovered after our definition was published in a conference paper. So we updated the definition for the journal version of the paper. Later work showed that the problem that caused us to change the definition was not as serious as originally thought. However, recently I have considered a definition that is much simpler than either of the two previous versions, which has the additional merit of dealing better with a number of problems. It is not clear that the third definition is the last word. Moreover, showing how the earlier definitions deal with the various problems that have been posed lends further insight into the difficulties and subtleties involved with finding a definition of causality. Thus, I consider all three definitions in this book. Following Einstein, my goal is to make things as simple as possible, but no simpler!
Like the definition of truth on which it depends, the HP definition of causality is relative to a causal setting. All three variants of the definition consist of three clauses. The first and third are simple and straightforward, and the same for all the definitions. All the work is done by the second clause; this is where the definitions differ.

**Definition 2.2.1** \( \vec{X} = \vec{x} \) is an actual cause of \( \varphi \) in the causal setting \((M, \vec{u})\) if the following three conditions hold:

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

**AC2.** See below.

**AC3.** \( \vec{X} \) is minimal; there is no strict subset \( \vec{X}' \) of \( \vec{X} \) such that \( \vec{X}' = \vec{x}' \) satisfies conditions AC1 and AC2, where \( \vec{x}' \) is the restriction of \( \vec{x} \) to the variables in \( \vec{X} \).

AC1 just says that \( \vec{X} = \vec{x} \) cannot be considered a cause of \( \varphi \) unless both \( \vec{X} = \vec{x} \) and \( \varphi \) actually happen. (I am implicitly taking \((M, \vec{u})\) to characterize the “actual world” here.) AC3 is a minimality condition, which ensures that only those elements of the conjunction \( \vec{X} = \vec{x} \) that are essential are considered part of a cause; inessential elements are pruned. Without AC3, if dropping a lit match qualified as a cause of the forest fire, then dropping a match and sneezing would also pass the tests of AC1 and AC2. AC3 serves here to strip “sneezing” and other irrelevant, over-specific details from the cause. AC3 can be viewed as capturing part of the spirit of the INUS condition; roughly speaking, it says that all the primitive events in the cause are necessary for the effect.

It is now time to bite the bullet and look at AC2. In the first two versions of the HP definition, AC2 consists of two parts. I start by considering the first of them, denoted AC2(a).

AC2(a) is a necessity condition. It says that for \( X = x \) to be a cause of \( \varphi \), there must be a value \( x' \) in the range of \( X \) such that if \( X \) is set to \( x' \), \( \varphi \) no longer holds. This is the but-for clause; but for the fact that \( X = x \) occurred, \( \varphi \) would not have occurred. As we saw in the Billy-Suzy rock-throwing example in Chapter 1, the naive but-for clause will not suffice. We must be allowed to apply it under certain contingencies, that is, under certain counterfactual settings, where some variables are set to values other than those they take in the actual situation or held to certain values while other variables change. For example, in the case of Suzy and Billy, we consider a contingency where Billy does not throw.

Here is the formal version of the necessity condition:

**AC2(a).** There is a partition of \( \mathcal{V} \) (the set of endogenous variables) into two disjoint subsets \( \vec{Z} \) and \( \vec{W} \) (so that \( \vec{Z} \cap \vec{W} = \emptyset \)) 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

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

Using statisticians’ notation, if \( \varphi \) is \( \vec{Y} = \vec{y} \), then the displayed formula becomes \( Y_{\vec{x}', \vec{w}}(\vec{u}) \neq \vec{y} \).

Roughly speaking, AC2(a) says that the but-for condition holds under the contingency \( \vec{W} = \vec{w} \). We can think of the variables in \( \vec{Z} \) as making up the “causal path” from \( \vec{X} \) to \( \varphi \). Intuitively, changing the value of some variable in \( \vec{X} \) results in changing the value(s) of some
variable(s) in $\vec{Z}$, which results in the values of some other variable(s) in $\vec{Z}$ being changed, which finally results in the truth value of $\varphi$ changing. (Whether we can think of the variables in $\vec{Z}$ as making up a causal path from some variable in $\vec{X}$ to some variable in $\varphi$ turns out to depend in part on which variant of the HP definition we consider. See Section 2.9 for a formalization and more discussion.) 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.

Unfortunately, AC1, AC2(a), and AC3 do not suffice for a good definition of causality. In the rock-throwing example, with just AC1, AC2(a), and AC3, Billy would be a cause of the bottle shattering. We need a sufficiency condition to block Billy. The sufficiency condition formalized by the following condition:

\textbf{AC2(b$^o$). If $\vec{z}^*$ is such that $(M, \bar{u}) \models \vec{Z} = \vec{z}^*$, then for all subsets $\vec{Z}'$ of $\vec{Z} - \vec{X}$, we have}

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

Again, taking $\varphi$ to be $\vec{Y} = \vec{y}$, in statisticians’ notation this becomes “if $\vec{Z}(\bar{u}) = \vec{z}^*$, then for all subsets $\vec{Z}'$ of $\vec{Z}$, we have $\vec{Y}_{\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*}(\bar{u}) = \vec{y}$.” It is important to write at least the $\vec{Z}'$ explicitly in the subscript here, since we are quantifying over it.

Note that, due to setting $\vec{W}$ to $\vec{w}$, the values of the variables in $\vec{Z}$ may change. AC2(b$^o$) says that this change does not affect $\varphi$; $\varphi$ continues to be true. Indeed, $\varphi$ continues to be true even if some variables in $\vec{Z}$ are forced to their original values.

Before going on, I need to make a brief technical digression to explain a slight abuse of notation in AC2(b$^o$). Suppose that $\vec{Z} = (Z_1, Z_2)$, $\vec{z} = (1, 0)$, and $\vec{Z}' = (Z_1)$. Then $\vec{Z}' \leftarrow \vec{z}$ is intended to be an abbreviation for $Z_1 \leftarrow 1$; that is, I am ignoring the second component of $\vec{z}$ here. More generally, when I write $\vec{Z}' \leftarrow \vec{z}$, I am picking out the values in $\vec{z}$ that correspond to the variables in $\vec{Z}'$ and ignoring those that correspond to the variables in $\vec{Z} - \vec{Z}'$. I similarly write $\vec{W}' \leftarrow \vec{w}$ if $\vec{W}'$ is a subset of $\vec{W}$.

For reasons that will become clearer in Example 2.8.1, the original definition was updated to use a stronger version of AC2(b$^o$). Sufficiency is now required to hold if the variables in any subset $\vec{W}'$ of $\vec{W}$ are set to the values in $\vec{w}$ (as well as the variables in any subset $\vec{Z}'$ of $\vec{Z} - \vec{X}$ being set to their values in the actual context). Taking $\vec{Z}$ and $\vec{W}$ as in AC2(a) (and AC2(b$^o$), here is the updated definition:

\textbf{AC2(b$^u$). If $\vec{z}^*$ is such that $(M, \bar{u}) \models \vec{Z} = \vec{z}^*$, then for all subsets $\vec{W}'$ of $\vec{W}$ and subsets $\vec{Z}'$ of $\vec{Z} - \vec{X}$, we have $(M, \bar{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*] \varphi.$}

In statisticians’ notation, this becomes “if $\vec{Z}(\bar{u}) = \vec{z}^*$, then for all subsets $\vec{Z}'$ of $\vec{Z}$ and $\vec{W}'$ of $\vec{W}$, we have $\vec{Y}_{\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*}(\bar{u}) = \vec{y}$.” Again, we need to write the variables in the subscript here.

The only difference between AC2(b$^o$) and AC2(b$^u$) lies in the clause “for all subsets $\vec{W}'$ of $\vec{W}$: AC2(b$^u$) must hold even if only a subset $\vec{W}'$ of the variables in $\vec{W}$ are set to their
2.2 A Formal Definition of Actual Cause

values in $\vec{w}$. This means that the variables in $\vec{W} - \vec{W}'$ essentially act as they do in the real world; that is, their values are determined by the structural equations, rather than being set to their values in $\vec{w}$.

The superscripts $o$ and $u$ in AC2(b$^o$) and AC2(b$^u$) are intended to denote “original” and “updated”.

I conclude by considering the simpler modified definition. This definition is motivated by the intuition that only what happens in the actual situation should matter. Thus, the only settings of variables allowed are ones that occur in the actual situation. Specifically, the modified definition simplifies AC2(a) by requiring that the only setting $\vec{w}$ of the variables in $\vec{W}$ that can be considered is the value of these variables in the actual context. Here is the modified AC2(a), denoted AC2(a$^m$) (the $m$ stands for “modified”):

AC2(a$^m$). There is a set $\vec{W}$ of variables in $\mathcal{V}$ and a setting $\vec{x'}$ of the variables in $\vec{X}$ such that if $(M, \vec{u}) \models \vec{W} = \vec{w}^*$, then

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

That is, we can show the counterfactual dependence of $\varphi$ on $\vec{X}$ by keeping the variables in $\vec{W}$ fixed at their actual values. In statisticians’ notation (with $\varphi$ being $\vec{Y} = \vec{y}$), this becomes “if $\vec{W}(\vec{u}) = \vec{w}^*$, then $\vec{Y}(\vec{u}) \neq \vec{y}^*$.”

It is easy to see that AC2(b$^o$) holds if all the variables in $\vec{W}$ are fixed at their values in the actual context: because $\vec{w}^*$ records the value of the variables in $\vec{W}$ in the actual context, if $\vec{X}$ is (re)set to $\vec{x}$, its value in the actual context, then $\varphi$ must hold (since, by AC1, $\vec{X} = \vec{x}$ and $\varphi$ both hold in the actual context); that is, we have $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}^*] \varphi$ if $\vec{w}^*$ is the value of the variables in $\vec{W}$ in the actual context. For similar reasons, AC2(b$^u$) holds if all the variables in $\vec{W}$ are fixed at their values in the actual context. Thus, there is no need for an analogue to AC2(b$^o$) or AC2(b$^u$) in the modified definition. This shows that the need for a sufficiency condition arises only if we are considering contingencies that differ from the actual setting in AC2(a). It is also worth noting that the modified definition does not need to mention $\vec{Z}$ (although $\vec{Z}$ can be taken to be the complement of $\vec{W}$).

For future reference, for all variants of the HP definition, the tuple $(\vec{W}, \vec{u}, \vec{x'})$ in AC2 is said to be a witness to the fact that $\vec{X} = \vec{x}$ is a cause of $\varphi$. (I take the witness to be $(\emptyset, \emptyset, \vec{x'})$ in the special case that $\vec{W} = \emptyset$.)

As I said, AC3 is a minimality condition. Technically, just as there are three versions of AC2, there are three corresponding versions of AC3. For example, in the case of the modified definition, AC3 should really say “there is no subset of $\vec{X}$ satisfying AC1 and AC2(a$^m$)”. I will not bother writing out these versions of AC3; I hope that the intent is clear whenever I refer to AC3.

If I need to specify which variant of the HP definition I am considering, I will say that $\vec{X} = \vec{x}$ is a cause of $\varphi$ according to the original (resp., updated, modified) HP definition. If $\vec{X} = \vec{x}$ is a cause of $\varphi$ according to all three variants, I often just say “$\vec{X} = \vec{x}$ is a cause of $\varphi$ in $(M, \vec{u})$”. Each conjunct in $\vec{X} = \vec{x}$ is called part of a cause of $\varphi$ in context $(M, \vec{u})$. As we shall see, what we think of as causes in natural language correspond to parts of causes, especially with the modified HP definition. Indeed, it may be better to use a term such as
Chapter 2. The HP Definition of Causality

“complete cause” for what I have been calling cause and then reserve “cause” for what I have called “part of a cause”. I return to this point after looking at a few examples.

Actually, although \( \vec{X} = \vec{x} \) is an abbreviation for the conjunction \( X_1 = x_1 \land \ldots \land X_k = x_k \), when we talk about \( \vec{X} = \vec{x} \) being a cause of \( \varphi \) (particularly for the modified HP definition), it might in some ways be better to think of the disjunction. Roughly speaking, \( \vec{X} = \vec{x} \) is a cause of \( \varphi \) if it is the case that if none of the variables in \( \vec{X} \) had their actual value, then \( \varphi \) might not have occurred. However, if \( \vec{X} \) has more than one variable, then just changing the value of a single variable in \( \vec{X} \) (or, indeed, any strict subset of the variables in \( \vec{X} \)) is not enough by itself to bring about \( \neg \varphi \) (given the context). Thus, there is a sense in which the disjunction can be viewed as a but-for cause of \( \varphi \). This will be clearer when we see the examples.

Note that all three variants of the HP definition declare \( X = x \) a cause of itself in \((M, u)\) as long as \((M, u) \models X = x \). This does not seem to me so unreasonable. At any rate, it seems fairly harmless, so I do not bother excluding it (although nothing would change if we did exclude it).

At this point, ideally, I would prove a theorem showing that some variant of the HP definition of actual causality is the “right” definition of actual causality. But I know of no way to argue convincingly that a definition is the “right” one; the best we can hope to do is to show that it is useful. As a first step, I show that all the definitions agree in the simplest and arguably most common case: but-for causes. Formally, say that \( X = x \) is a but-for cause of \( \varphi \) in \((M, \vec{u})\) if AC1 holds (so that \((M, \vec{u}) \models (X = x) \land \varphi \)) and there exists some \( x' \) such that \((M, \vec{u}) \models [X \leftarrow x'] \neg \varphi \). Note here I am assuming that the cause is a single conjunct.

**Proposition 2.2.2** If \( X = x \) is a but-for cause of \( Y = y \) in \((M, \vec{u})\), then \( X = x \) is a cause of \( Y = y \) according to all three variants of the HP definition.

**Proof:** Suppose that \( X = x \) is a but-for cause of \( Y = y \). There must be a possible value \( x' \) of \( X \) such that \((M, \vec{u}) \models [X \leftarrow x'] \neg \varphi \). Then \((\emptyset, \emptyset, x') \) (i.e., \( \vec{W} = \emptyset \) and \( X = x' \)) is a witness to \( X = x \) being a cause of \( \varphi \) for all three variants of the definition. Thus, AC2(a) and AC2(a") hold if we take \( \vec{W} = \emptyset \). Since \((M, \vec{u}) \models X = x \), if \((M, \vec{u}) \models \vec{Z} = \vec{z}^* \), where \( \vec{Z} = \vec{V} - \{X\} \), then it is easy to see that \((M, \vec{u}) \models [X \leftarrow x](\vec{Z} = \vec{z}^*) \): since \( X = x \) and \( \vec{Z} = \vec{z}^* \) in the (unique) solution to the equations in context \( \vec{u} \), this is still the case in the unique solution to the equations in \( M_{X \leftarrow x} \). For similar reasons, \((M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{Z}' \leftarrow \vec{z}^*] \varphi \) for all subsets \( \vec{Z}' \) of \( \vec{V} - \{X\} \). (See Lemma 2.10.2, where these statements are formalized.) Thus, AC2(b") holds. Because \( \vec{W} = \emptyset \), AC2(b") follows immediately from AC2(b").

Of course, the definitions do not always agree. Other connections between the definitions are given in the following theorem. In the statement of the theorem, the notation \( |\vec{X}| \) denotes the cardinality of \( \vec{X} \).
Theorem 2.2.3

(a) If $X = x$ is part of a cause of $\varphi$ in $(M, \vec{u})$ according to the modified HP definition, then $X = x$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the original HP definition.

(b) If $X = x$ is part of a cause of $\varphi$ in $(M, \vec{u})$ according to the modified HP definition, then $X = x$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the updated HP definition.

(c) If $X = x$ is part of a cause of $\varphi$ in $(M, \vec{u})$ according to the updated HP definition, then $X = x$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the original HP definition.

(d) If $\vec{X} = \vec{x}$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the original HP definition, then $|\vec{X}| = 1$ (i.e., $\vec{X}$ is a singleton).

Of course, (a) follows from (b) and (c); I state it separately just to emphasize the relations. Although it may not look like it at first glance, part (d) is quite similar in spirit to parts (a), (b), and (c); an equivalent reformulation of (d) is: “If $X = x$ is part of a cause of $\varphi$ in $(M, \vec{u})$ according to the original HP definition, then $X = x$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the original HP definition.” Call this statement (d'). Clearly (d) implies (d'). For the converse, suppose that (d') holds, $\vec{X} = \vec{x}$ is a cause of $\varphi$ in $(M, \vec{u})$, and $|\vec{X}| > 1$. Then if $X = x$ is a conjunct of $\vec{X} = \vec{x}$, by (d'), $X = x$ is a cause of $\varphi$ in $(M, \vec{u})$. But by AC3, this can hold only if $X = x$ is the only conjunct of $\vec{X} = \vec{x}$, so $|\vec{X}| = 1$.

Parts (a), (b), and (c) of Theorem 2.2.3 show that, in a sense, the original HP definition is the most permissive of the three and the modified HP definition is the most restrictive, with the updated HP definition lying somewhere in between. The converses of (a), (b), and (c) do not hold. As Example 2.8.1 shows, if $X = x$ is a cause of $\varphi$ according to the original HP definition, it is not, in general, part of a cause according to the modified or updated HP definition. In this example, we actually do not want causality to hold, so the example can be viewed as showing that the original definition is too permissive (although this claim is not quite as clear as it may appear at first; see the discussion in Section 2.8). Example 2.8.2 shows that part of a cause according to the updated HP definition need not be (part of) a cause according to the modified HP definition.

The bottom line here is that, although these definitions often agree, they do not always agree. It is typically on the most problematic examples on which they disagree. This is perhaps not surprising given that the original definition was updated and modified precisely to handle such examples. However, even in cases where the original definition seems to give “wrong” answers, there are often ways of dealing with the problem that do not require modifying the definition. Moreover, the fact that causes are always singletons with the original definition makes it attractive in some ways. The jury is still out on what the “right” definition of causality is. Although my current preference is the modified HP definition, I will consider all three definitions throughout the book. Despite the extra burden for the reader (for which I apologize in advance!), I think doing so gives a deeper insight into the subtleties of causality.

The proof of Theorem 2.2.3 can be found in Section 2.10.1. I recommend that the interested reader defer reading the proof until after going through the examples in the next section, since they provide more intuition for the definitions.
2.3 Examples

Because I cannot prove a theorem showing that (some variant of) the HP definition is the “right” definition of causality, all I can do to argue that the definition is reasonable is to show how it works in some of the examples that have proved difficult for other definitions to handle. That is one of the goals of this section. I also give examples that illustrate the subtle differences between the variants of the HP definition. I start with a few basic examples mainly intended to show how the definition works.

Example 2.3.1 For the forest-fire example, I consider two causal models, $M^c$ and $M^d$, for the conjunctive and disjunctive cases, respectively. These models were described earlier; the endogenous variables are $L$, $MD$, and $FF$, and $U$ is the only exogenous variable. In both cases, I want to consider the context $(1, 1)$, so the lightning strikes and the arsonist drops the match. In the conjunctive model, both the lightning and the dropped match are but-for causes of the forest fire; if either one had not occurred, the fire would not have happened. Hence, by Proposition 2.2.2, both $L = 1$ and $MD = 1$ are causes of $FF = 1$ in $(M^c, (1, 1))$ according to all three variants of the definition. By AC3, it follows that $L = 1 \land MD = 1$ is not a cause of $FF = 1$ in $(M^c, (1, 1))$. This already shows that causes might not be unique; there may be more than one cause of a given outcome.

However, in the disjunctive case, there are differences. With the original and updated definition, again, we have that both $L = 1$ and $MD = 1$ are causes of $FF = 1$ in $(M^d, (1, 1))$. I give the argument in the case of $L = 1$; the argument for $MD = 1$ is identical. Clearly $(M^d, (1, 1)) \models FF = 1$ and $(M^d, (1, 1)) \models L = 1$; in the context $(1, 1)$, the lightning strikes and the forest burns down. Thus, AC1 is satisfied. AC3 is trivially satisfied: since $\vec{X}$ consists of only one element, $L$, $\vec{X}$ must be minimal.

For AC2, as suggested earlier, let $\vec{Z} = \{L, FF\}$, $\vec{W} = \{MD\}$, $x' = 0$, and $w = 0$. Clearly, $(M^d, (1, 1)) \models [L \leftarrow 0, MD \leftarrow 0](FF \neq 1)$; if the lightning does not strike and the match is not dropped, the forest does not burn down, so AC2(a) is satisfied. To see the effect of the lightning, we must consider the contingency where the match is not dropped; AC2(a) allows us to do that by setting $MD$ to 0. (Note that setting $L$ and $MD$ to 0 overrides the effects of $U$; this is critical.) Moreover,

$$(M^d, (1, 1)) \models [L \leftarrow 1, MD \leftarrow 0](FF = 1) \text{ and } (M^d, (1, 1)) \models [L \leftarrow 1](FF = 1);$$

in context $(1, 1)$, if the lightning strikes, then the forest burns down even if the lit match is not dropped, so AC2(b$^o$) and AC2(b$^u$) are satisfied. (Note that since $\vec{Z} = \{L, FF\}$, the only subsets of $\vec{Z} - \vec{X}$ are the empty set and the singleton set consisting of just $FF$; similarly, since $\vec{W} = \{MD\}$, the only subsets of $\vec{W}$ are the empty set and the singleton set consisting of $MD$ itself. Thus, we have considered all the relevant cases here.)

This argument fails in the case of the modified definition because the setting $MD = 0$ is not allowed. The only witnesses that can be considered are ones where $\vec{W}$ has the value it does in the actual context; in the actual context here, $MD = 1$. So, with the modified definition, neither $L = 1$ nor $MD = 1$ is a cause. However, it is not hard to see that $L = 1 \land MD = 1$ is a cause of $FF = 1$. This shows why it is critical to consider only single conjuncts in Proposition 2.2.2 and Theorem 2.2.3 is worded in terms of parts of causes. Although $L =$
2.3 Examples

$1 \land MD = 1$ is a cause of $FF = 1$ according to the modified HP definition, it is not a cause according to either the original or updated HP definition. In contrast, $L = 1$ and $MD = 1$ are both parts of a cause of $FF = 1$ according to all three definitions.

It is arguably a feature of the original and updated HP definition that they call both $L = 1$ and $MD = 1$ causes of $FF = 1$. Calling the conjunction $L = 1 \land MD = 1$ a cause of $FF = 1$ does not seem to accord with natural language usage. There are two ways to address this concern. As I suggested earlier, with the modified HP definition, it may be better to think of parts of causes as coming closer to what we call causes in natural language. That would already deal with this concern. A second approach, which I also hinted at earlier, is to observe that it may be better to think of the disjunction $L = 1 \lor MD = 1$ as being the cause. Indeed, we can think of $MD = 1 \lor L = 1$ as a but-for cause of $FF = 1$; if it does not hold, then it must be the case both $L = 0$ and $MD = 0$, so there is no fire. The reader should keep in mind both of these ways of thinking of conjunctive causes with the modified HP definition.

As we shall see, the notion of responsibility, discussed in Chapter 6, allows the original and updated HP definition to distinguish these two scenarios, as does the notion of sufficient cause, discussed in Section 2.6.

This simple example already reveals some of the power of the HP definition. The case where the forest fire is caused by either the lightning or the dropped match (this is said to be a case of overdetermination in the literature) cannot be handled by the simple but-for definition used in the law. It seems reasonable to call both the lightning and the dropped match causes, or at least parts of causes. We certainly wouldn’t want to say that there is no cause, as a naive application of the but-for definition would do. (However, as I noted above, if we allow disjunctive causes, then a case can be made that the disjunction $L = 1 \lor MD = 1$ is a but-for cause of the forest fire.) Overdetermination occurs frequently in legal cases; a victim may be shot by two people, for example. It occurs in voting as well; I look at this a little more carefully in the following example.

Example 2.3.2 Consider the voting scenario discussed earlier where there are 11 voters. If Suzy wins 6–5, then all the definitions agree that each of the voters for Suzy is a cause of Suzy’s victory; indeed, they are all but-for causes. But suppose that Suzy wins 11–0. Here we have overdetermination. The original and updated HP definition would still call each of the voters a cause of Suzy winning; the witness involves switching the votes of 5 voters. Since the modified HP definition does not allow such switching, according to the modified HP definition, any subset of six voters is a cause of Suzy winning (and every individual is part of a cause). Again, if we think of the subset as being represented by a disjunction, it can be thought of as a but-for cause of Suzy winning. If all six voters had switched their votes to Billy, then Suzy would not have won. Minimality holds; if a set of fewer than six voters had voted for Billy, then Suzy would still have won.

We do have an intuition that a voter for Suzy is somehow “less” of a cause of the victory in an 11–0 victory than in a 6–5 victory. This intuition is perhaps most compatible with the modified HP definition. In this case, $\vec{X} = \vec{x}$ is a cause of $\varphi$ if $\vec{X}$ is a minimal set of variables whose values have to change for $\neg \varphi$ to hold. The bigger $\vec{X}$ is, the less of a cause each of its conjuncts is. I return to this issue when I discuss responsibility in Chapter 6.
Example 2.3.3 Now I (finally!) consider the rock-throwing example, where Suzy and Billy both throw rocks at a bottle, but Suzy’s hits the bottle, and Billy’s doesn’t (although it would have hit had Suzy’s not hit first). We get the desired result—that Suzy’s throw is a cause, but Billy’s is not—but only if we model the story appropriately. Consider first a coarse causal model, with three endogenous variables:

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

For simplicity, assume that there is one exogenous variable $u$, which determines whether Billy and Suzy throw. (In most of the subsequent examples, I omit the exogenous variable in both the description of the story and the corresponding causal network.) Take the formula for $BS$ to be such that the bottle shatters if either Billy or Suzy throws; that is, $BS = BT \lor ST$. (I am implicitly assuming that Suzy and Billy never miss if they throw. Also, I follow the fairly standard mathematics convention of saying “if” rather than “if and only if” in definitions; clearly, the equation is saying that the bottle shatters if and only if either Billy or Suzy throws.) For future reference, I call this model $M_{RT}$ (where $RT$ stands for “rock throwing”).

$BT$ and $ST$ play symmetric roles in $M_{RT}$; there is nothing to distinguish them. Not surprisingly, both Billy’s throw and Suzy’s throw are classified as causes of the bottle shattering in $M_{RT}$ in the context $u$ where Suzy and Billy both throw according to the original and updated HP definition (the conjunction $ST = 1 \land BT = 1$ is the cause according to the modified HP definition). The argument is essentially identical to that used for the disjunctive model of the forest-fire example, where either the lightning or the dropped match is enough to start the fire. Indeed, the causal network describing this situation looks like that in Figure 2.1, with $ST$ and $BT$ replacing $L$ and $MD$. For convenience, I redraw the network here. As I said earlier, here and in later figures, I typically omit the exogenous variable(s).

![Figure 2.2: $M_{RT}$](image)

The trouble with $M_{RT}$ 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. $M_{RT}$ has to be refined to express this distinction. One way is to invoke a dynamic model. Although this can be done (see the notes at the end of the chapter for more discussion), a simpler way to gain the expressiveness needed here 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). I leave it to the reader to check that $ST = 1$ is a cause...
of $BS = 1$, but $BT = 1$ is not (if Suzy doesn’t throw but Billy does, then we would have $BS = 2$). 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” in the definition of $BS = 2$, which requires the identification of the actual cause.

A more useful choice is to add two new variables to the model:

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

We now modify the equations as follows:

- $BS = 1$ if $SH = 1$ or $BH = 1$;
- $SH = 1$ if $ST = 1$;
- $BH = 1$ if $BT = 1$ and $SH = 0$.

Thus, Billy’s throw hits if Billy throws and Suzy’s rock doesn’t hit. The last equation implicitly assumes that Suzy throws slightly ahead of Billy, or slightly harder.

Call this model $M'_{RT}$. $M'_{RT}$ is described by the graph in Figure 2.3 (where again the exogenous variables are omitted). The asymmetry between $BH$ and $SH$ (in particular, the fact that Billy’s throw doesn’t hit the bottle if Suzy throws) is modeled by the fact that there is an edge from $SH$ to $BH$ but not one in the other direction; $BH$ depends (in part) on $SH$ but not vice versa.

![Figure 2.3: $M'_{RT}$—a better model for the rock-throwing example.](http://direct.mit.edu/books/book/chapter-pdf/269281/9780262336611_cab.pdf)

Taking $u$ to be the context where Billy and Suzy both throw, according to all three variants of the HP definition, $ST = 1$ is a cause of $BS = 1$ in $(M'_{RT}, u)$, but $BT = 1$ is not. To see that $ST = 1$ is a cause according to the original and updated HP definition, note that it is immediate that AC1 and AC3 hold. To see that AC2 holds, one possibility is to choose $\bar{Z} = \{ST, SH, BH, BS\}$, $\bar{W} = \{BT\}$, and $w = 0$. When $BT$ is set to 0, $BS$ tracks $ST$: if Suzy throws, the bottle shatters, and if she doesn’t throw, the bottle does not shatter. It immediately follows that AC2(a) and AC2(b°) hold. AC2(b°) is equivalent to AC2(b°) in this case, since $\bar{W}$ is a singleton.
To see that $BT = 1$ is not a cause of $BS = 1$ in $(M'_{RT}, u)$, we must check that there is no partition of the endogenous variables into sets $\vec{Z}$ and $\vec{W}$ that satisfies AC2. Attempting the symmetric choice with $\vec{Z} = \{BT, BH, SH BS\}$, $\vec{W} = \{ST\}$, and $w = 0$ violates AC2(b$^o$) and AC2(b$^u$). To see this, take $\vec{Z}' = \{BH\}$. In the context where Suzy and Billy both throw, $BH = 0$. If $BH$ is set to 0, then the bottle does not shatter if Billy throws and Suzy does not. It is precisely because, in this context, Suzy’s throw hits the bottle and Billy’s does not that we declare Suzy’s throw to be the cause of the bottle shattering. AC2(b$^o$) and AC2(b$^u$) capture that intuition by forcing us to consider the contingency where $BH = 0$ (i.e., where $BH$ takes on its actual value), despite the fact that Billy throws.

Of course, just checking that this particular partition into $\vec{Z}$ and $\vec{W}$ does not make $BT = 1$ a cause does not suffice to establish that $BT = 1$ is not a cause according to the original or updated definitions; we must check all possible partitions. I just sketch the argument here: The key is to consider whether $BH$ is in $\vec{W}$ or $\vec{Z}$. If $BH$ is in $\vec{W}$, then how we set $BT$ has no effect on the value of $BS$: If $BH$ is set to 0, then the value of $BS$ is determined by the value of $ST$; and if $BH$ is set to 1, then $BS = 1$ no matter what $BT$ is. (It may seem strange to intervene and set $BH$ to 1 if $BT = 0$. It means that somehow Billy hits the bottle even if he doesn’t throw. Such “miraculous” interventions are allowed by the definition; part of the motivation for the notion of normality considered in Chapter 3 is to minimize their usage.) This shows that we cannot have $BH$ in $\vec{W}$. If $BH$ is in $\vec{Z}$, then we get the same problem with AC2(b$^o$) or AC2(b$^u$) as above; it is easy to see that at least one of $SH$ or $ST$ must be in $\vec{W}$, and $\vec{w}$ must be such that whichever is in $\vec{W}$ is set to 0.

Note that, in this argument, it is critical that in AC2(b$^o$) and AC2(b$^u$) we allow setting an arbitrary subset of variables in $\vec{Z} - \vec{X}$ to their original values. There is a trivial reason for this: if we set all variables in $\vec{Z} - \vec{X}$ to their original values in $M'_{RT}$, then, among other things, we will set $BS$ to 1. We will never be able to show that Billy’s throw is not a cause if $BS$ is set to 1. But even requiring that all the variables in $\vec{Z} - \{BS, BT\}$ be set to their original values does not work. For suppose that we take $\vec{W} = \{ST\}$ and take $\vec{w}$ such that $ST = 0$. Clearly AC2(a) holds because if $BT = 0$, then $BS = 0$. But if all the variables in $\vec{Z} - \{BT, BS\}$ are set to their original values, then $SH$ is set to 1, so the bottle shatters. To show that $BT = 1$ is not a cause, we must be able to set $BH$ to its original value of 0 while keeping $SH$ at 0. Setting $BH$ to 0 captures the intuition that Billy’s throw is not a cause because, in the actual world, his rock did not hit the bottle ($BH = 0$).

Finally, consider the modified HP definition. Here things are much simpler. In this case, taking $\vec{W} = \{BT\}$ does not work to show that $ST = 1$ is a cause of $BS = 1$; we are not allowed to change $BT$ from 1 to 0. However, we can take $\vec{W} = \{BH\}$. Fixing $BH$ at 0 (its setting in the actual context), if $ST$ is set to 0, then $BS = 0$, that is, $(M, u) \models [ST \leftarrow 0, BH \leftarrow 0] (BS = 0)$. Thus, $ST = 1$ is a cause of $BS = 1$ according to the modified HP definition. (Taking $\vec{W} = \{BH\}$ would also have worked to show that $ST = 1$ is a cause of $BS = 1$ according to the modified definition.) Indeed, by Theorem 2.2.3, showing that $ST = 1$ is a cause of $BS = 1$ according to the modified definition suffices to show that it is also a cause according to the original and updated definition. I went through the argument for the original and updated definition because the more obvious choice for $\vec{W}$ in that case is $\{BT\}$, and it works under those definitions.) But now it is also easy to check that $BT = 1$ is not a cause of $BS = 1$ according to the modified HP definition. No matter
which subset of variables other than $BT$ are held to their values in $u$ and no matter how we set $BT$, we have $SH = 1$ and thus $BS = 1$.

I did not completely specify $M'_{RT}$ in the example above. Specifically, I did not specify the set of possible contexts. What happens in other contexts is irrelevant for determining whether $ST = 1$ or $BT = 1$ is a cause of $BS = 1$ in $(M'_{RT}, u)$. But once I take normality considerations into account in Chapter 3, it will become relevant. I described $u$ above as the context where Billy and Suzy both throw, implicitly suggesting that there are four contexts, determining the four possible combinations of Billy throwing and Suzy throwing. So we can take $M_{RT}'$ to be the model with these four contexts.

There is arguably a problem with this definition of $M_{RT}'$: it “bakes in” the temporal ordering of events, in particular, that Suzy’s rock hits before Billy’s rock. If we want a model that describes “rock-throwing situations for Billy and Suzy” more generally, we would not want to necessarily assume that Billy and Suzy are accurate, nor that Suzy always hits before Billy if they both throw. Whatever assumptions we make along these lines can be captured by the context. That is, we can have the context determine (a) whether Suzy and Billy throw, (b) how accurate they are (would Billy hit if he were the only one to throw, and similarly for Suzy), and (c) whose rock hits first if they both throw and are accurate (allowing for the possibility that they both hit at the same time). This richer model would thus have 48 contexts: there are four choices of which subset of Billy and Suzy throw, four possibilities for accuracy (both are accurate, Suzy is accurate and Billy is not, and so on), and three choices regarding who hits first if they both throw. If we use this richer model, we would also want to expand the language to include variables such as $SA$ (Suzy is accurate), $BA$ (Billy is accurate), $SF$ (Suzy’s throw arrives first if both Billy and Suzy throw), and $BF$ (Billy’s throw arrives first). Call this richer model $M_{RT}^*$. In $M_{RT}^*$, the values of $ST$, $BT$, $SA$, $BA$, $SF$, and $BF$ are determined by the context. As with $M_{RT}$ and $M'_{RT}$, the bottle shatters if either Suzy or Billy hit it, so we have the equation $BS = BH \lor SH$. But now the equations for $BH$ and $SH$ depend on $ST$, $BT$, $SA$, $BA$, $SF$, and $BF$. Of course $SH = 0$ if Suzy doesn’t throw (i.e., if $ST = 0$) or is inaccurate ($SA = 0$), but even in this case, the equations need to tell us what would have happened had Suzy thrown and been accurate, so that we can determine the truth of formulas such as $[ST \leftarrow 1, SA \leftarrow 1](SH = 0)$. As suggested by the discussion above, the equation for $SH$ is

$$SH = \begin{cases} 1 & \text{if } ST = 1, SA = 1, \text{and either } BT = 0, BA = 0, \text{or } SF = 1; \\ 0 & \text{otherwise.} \end{cases}$$

In words, Suzy hits the bottle if she throws and is accurate, and either (a) Billy doesn’t throw, (b) Billy is inaccurate, or (c) Billy throws and is accurate but Suzy’s rock arrives at the bottle first (or at the same time as Billy’s); otherwise, Suzy doesn’t hit the bottle.

The context $u$ in $M'_{RT}$ corresponds to the context $u^*$ in $M_{RT}^*$ where Billy and Suzy both throw, both are accurate, and Suzy’s throw hits first. Exactly the same argument as that above shows that $ST = 1$ is a cause of $(M_{RT}^*, u^*)$, and $BT = 1$ is not. In the model $M_{RT}^*$, the direction of the edge in the causal network between $SH$ and $BH$ depends on the context. In some contexts (the ones where both Suzy and Billy are accurate and Suzy would hit first if they both throw), $BH$ depends on $SH$; in other contexts, $SH$ depends on $BH$. As a result, $M_{RT}^*$ is not what I called a strongly recursive model. $M_{RT}^*$ is still a recursive model because in each context $u'$ in $M_{RT}^*$, there are no cyclic dependencies; there is still an ordering $\leq_{u'}$ on
the endogenous variables such that unless \( X \preceq_{u'} Y, \) \( Y \) is independent of \( X \) in \((M'_{RT}, u')\). In particular, although \( SH \preceq_{u^*} BH \) and there is a context \( u'' \) in \( M'_{RT} \) such that \( BH \preceq_{u''} SH, \) \( M'_{RT} \) is still considered recursive.

It is critical in this analysis of the rock-throwing example that the model includes the variables \( BH \) and \( SH \), that is, that Billy’s rock hitting the bottle is taken to be a different event than Suzy’s rock hitting the bottle (although there is no problem if it includes extra variables). To understand the need for \( BH \) and \( SH \) (or some analogous variables), consider an observer watching the situation. Why would she declare Suzy’s throw to be a cause and Billy’s throw not to be a cause? Presumably, precisely because it was Suzy’s rock that hit the bottle and not Billy’s. If this is the reason for the declaration, then it must be modeled. Of course, using the variables \( BH \) and \( SH \) is not the only way of doing this. All that is needed is that the language be rich enough to allow us to distinguish the situation where Suzy’s rock hits and Billy’s rock does not from the one where Billy’s rock hits and Suzy’s rock does not. (That is why I said “or some analogous variables” above.) An alternative approach to incorporating temporal information is to have time-indexed variables (e.g., to have a family of variables \( BS_k \) for “bottle shatters at time \( k \)” and a family \( H_k \) for “bottle is hit at time \( k \)”).

With these variables and the appropriate equations, we can dispense with \( SH \) and \( BH \) and just have the variables \( H_1, H_2, \ldots \) that talk about the bottle being hit at time \( k \), without the variable specifying who hit the bottle. For example, if we assume that all the action happens at times 1, 2, and 3, then the equations are \( H_1 = ST \) (if Suzy throws, then the bottle is hit at time 1), \( H_2 = BT \land \neg H_1 \) (the bottle is hit at time 2 if Billy throws and the bottle was not already hit at time 1), and \( BS = H_1 \lor H_2 \) (if the bottle is hit, then it shatters). Again, these equations model the fact that Suzy hits first if she throws. And again, we could assume that the equations for \( H_1 \) and \( H_2 \) are context-dependent, so that Suzy hits first in \( u \), but Billy hits first in \( u' \). In any case, in context \( u \) where Suzy and Billy both throw but Suzy’s rock arrives at the bottle first, Suzy is still the cause, and Billy isn’t, using essentially the same analysis as that above.

To summarize the key point here, the language (i.e., the variables chosen and the equations for them) must be rich enough to capture the significant features of the story, but there is more than one language that can do this. As we shall see in Chapter 4, the choice of language can in general have a major impact on causality judgments.

The rock-throwing example emphasizes an important moral. If we want to argue in a case of preemption that \( X = x \) rather than \( Y = y \) is the cause of \( \varphi \), then there must be a 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. The need for such variables is certainly consistent with intuition and the way we present evidence. If we want to argue (say, in a court of law) that it was \( A \)'s shot that killed \( C \) and not \( B \)'s, then, assuming that \( A \) shot from \( C \)'s left and \( B \) from the right, we would present evidence such as the bullet entering \( C \) from the left side. The side from which the shot entered is the relevant variable in this case. The variable may involve temporal evidence (if \( C \)'s shot had been the lethal one, then the death would have occurred a few seconds later), but it certainly does not have to.

The rock-throwing example also emphasizes the critical role of what happens in the actual situation in the definition. Specifically, we make use of the fact that, in the actual context,
2.3 Examples

Billy’s rock did not hit the bottle. In the original and updated HP definition, this fact is applied in AC2(b): to show that Billy is a cause, we would have to show that the bottle shatters when Billy throws, even if BH is set to its actual value of 0. In the modified HP definition, it is applied in AC2(a") to show that Suzy’s throw is a cause: we can set ST = 0 while keeping BH fixed at 0. The analogous argument would not work to show that Billy’s throw is a cause.

Although what happens in the actual context certainly affects people’s causality judgments, why is the way that this is captured in AC2(a"), AC2(b"), or AC2(b") the “right” way of capturing it? I do not have a compelling answer to this question, beyond showing that these definitions work well in examples. I have tried to choose examples that, although not always realistic, capture important features of causality. For example, while the rock-throwing example may not seem so important in and of itself, it is an abstraction of a situation that arises frequently in legal cases, where one potential cause is preempted by another. Monopolistic practices by a big company cause a small company to go bankrupt, but it would have gone bankrupt anyway because of poor management; a smoker dies in a car accident, but he would have died soon due to inoperable lung cancer had there not been an accident. A good definition of causality is critical for teasing out what is a cause and what is not in such cases. Many of the other examples in this section are also intended to capture the essence of other important issues that arise in legal (and everyday) reasoning. It is thus worth understanding in all these cases exactly what the role of the actual context is.

Example 2.3.4  This example considers the problem of what has been called double prevention.

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, and pulls the trigger; Enemy’s plane goes down in flames. Suzy’s mission is undisturbed, and the bombing takes place as planned.

Is Billy a cause of the success of the mission? After all, he prevented Enemy from preventing Suzy from carrying out the mission. Intuitively, it seems that the answer is yes, and the obvious causal model gives us this. Suppose that we have the following variables:

- $BPT$ for “Billy pulls trigger”, with values 0 (he doesn’t) and 1 (he does);
- $EE$ for “Enemy eludes Billy”, with values 0 (he doesn’t) and 1 (he does);
- $ESS$ for “Enemy shoots Suzy”, with values 0 (he doesn’t) and 1 (he does);
- $SBT$ for “Suzy bombs 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 given in Figure 2.4.

In this model, $BPT = 1$ is a but-for cause of $TD = 1$, as is $SBT = 1$, so both $BPT = 1$ and $SBT = 1$ are causes of $TD = 1$ according to all the variants of the HP definition.
We can make the story more complicated by adding 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 \( EE = 1 \) (see Figure 2.5), but then, using the naive but-for criterion, one might conclude that the target will be destroyed \( (TD = 1) \) regardless of Billy’s action, so \( BPT = 1 \) would not be a cause of \( TD = 1 \). All three variants of the HP definition still declare \( BPT = 1 \) to be a cause of \( TD = 1 \), as expected. We now take \( \bar{W} = \{HPT\} \) and fix \( HPT \) at its value in the actual context, namely, 0. Although Billy’s action seems superfluous under ideal conditions, it becomes essential under a contingency where Hillary for some reason fails to pull the trigger. This contingency is represented by fixing \( HPT \) at 0 irrespective of \( EE \).

So far, this may all seem reasonable. But now go back to the original story and suppose that, again, Billy goes up along with Suzy, but Enemy does not actually show up. Is Billy going up a cause of the target being destroyed? Clearly, if Enemy had shown up, then Billy would have shot him down, and he would have been a but-for cause of the target being destroyed. But what about the context where Enemy does not show up? It seems that the original and updated HP definition would say that, even in that context, Billy going up is a cause of the target being destroyed. For if we consider the contingency where Enemy had shown up, then the target would not have been destroyed if Billy weren’t there, but since he is, the target is destroyed.

This may seem disconcerting. Suppose it is known that just about all the enemy’s aircraft have been destroyed, and no one has been coming up for days. Is Billy going up still a cause? The concern is that, if so, almost any \( A \) can become a cause of any \( B \) by telling a story of how there might have been a \( C \) that, but for \( A \), would have prevented \( B \) from happening (just as there might have been an Enemy that prevented the target being destroyed had Billy not been there).

Does the HP definition really declare Billy showing up a cause of the target being destroyed in this case? Let’s look a little more carefully at how we model this story. If we want to call
2.3 Examples

Billy going up a cause, then we need to add to the model a variable BGU corresponding to Billy going up. Similarly, we should add a variable ESU corresponding to Enemy showing up. We have the obvious equations, which say that (in the contexts of interest) BPT = 1 (i.e., Billy pulls the trigger) only if Billy goes up and Enemy shows up, and EE = 0 only if either Enemy doesn’t show up or Billy pulls the trigger. This gives us the model shown in Figure 2.6.

![Figure 2.6: Blowing up the target, where Enemy may not show up.](image)

In the actual world, where Billy goes up and Enemy doesn’t, Billy doesn’t pull the trigger (why should he?), so BPT = 0. Now it is easy to see that Billy going up (i.e., BGU = 1) is not a cause of the target being destroyed if Enemy doesn’t show up. AC2(b°) fails (and hence so does AC2(b°)): if Enemy shows up, the target won’t be destroyed, even if Billy shows up, because in the actual world BPT = 0. The modified HP definition addresses the problem even more straightforwardly. No matter which variables we fix to their actual values, TD = 1 even if BGU = 0.

Cases of prevention and double prevention are quite standard in the real world. It is standard to install fire alarms, which, of course, are intended to prevent fires. When the batteries in a fire alarm become weak, fire alarms typically “chirp”. Deactivating the chirping sound could be a cause of a fire not being detected due to double prevention: deactivating the chirping sound prevents the fire alarm from sounding, which in turn prevents the fire from being detected.

**Example 2.3.5** Can not performing an action be (part of) a cause? Consider the following story.

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 that the doctor does not treat Billy on Monday. Is the doctor’s not treating Billy a cause of Billy’s being sick on Tuesday? It seems that it should be, and indeed it is according to all variants of the HP definition. Suppose that icht is the context where, among other things, Billy is sick on Monday and the doctor forgets to administer the medication Monday. It seems reasonable that the model should have two variables:
Chapter 2. The HP Definition of Causality

- $MT$ 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 setting, $MT = 0$ is a but-for cause of $BMC = 1$, and hence a cause according to all three variants of the HP definition.

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 causal model above, the other doctors’ failure to give Billy his medication is not a cause, since the model has no variables to model the other doctors’ actions, just as there was no variable in the causal model of Example 2.3.1 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 variables corresponding to the other doctors, then they too would be causes of Billy’s being sick on Tuesday. The more refined definition of causality given in Chapter 3, which takes normality into account, provides a way of avoiding this problem even if the model includes endogenous variables for the other doctors.

Causation by omission is a major issue in the law. To take just one of many examples, a surgeon can be sued for the harm caused due to a surgical sponge that he did not remove after an operation.

The next example again emphasizes how the choice of model can change what counts as a cause.

**Example 2.3.6** 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 track instead of the left. Because the tracks reconverge up ahead, the train arrives at its destination all the same.

If we model this story using three 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 track) and 1 (it goes on the right track); and $A$ for “arrival”, with values 0 (the train does not arrive at the point of reconvergence) and 1 (it does)—then all three definitions agree that flipping the switch is not a cause of the train arriving.

But now suppose that we replace $T$ with two binary variables, $LB$ (for left track blocked), which is 0 if the left track is not blocked, and 1 if it is, and $RB$ (for right track blocked), defined symmetrically. Suppose that $LB$, $RB$, and $F$ are determined by the context, while the value of $A$ is determined in the obvious way by which track the train is going down (which is determined by how the switch is set) and whether the track that it is going down is blocked; specifically, $A = (F \land \neg LB) \lor (\neg F \land \neg RB)$. In the actual context, $F = 1$ and $LB = RB = 0$. Under the original and updated HP definition, $F = 1$ is a cause of $A = 1$. For in the contingency where $LB = 1$, if $F = 1$, then the train arrives, whereas if $F = 0$, then the train does not arrive. While adding the variables $LB$ and $RB$ suggests that we care about whether a track is blocked, it seems strange to call flipping the switch a cause of the train arriving when in fact both tracks are unblocked. This problem can be dealt with to some...
extent by invoking normality considerations, but not completely (see Example 3.4.3). With
the modified definition, the problem disappears. Flipping the switch is not a cause of the train
arriving if both tracks are unblocked, nor is it a cause of the train not arriving if both tracks
are blocked.

Example 2.3.7 Suppose that a captain and a sergeant stand before a private, both shout
“Charge!” at the same time, and the private charges. Some have argued that, because orders
from higher-ranking officers trump those of lower-ranking officers, the captain is a cause of
the charge, whereas the sergeant is not.

It turns out that which of the sergeant or captain is a cause depends in part on the variant
of the HP definition we consider and in part on what the possible actions of the sergeant and
captain are. First, suppose that the sergeant and the captain can each either order an advance,
order a retreat, or do nothing. Formally, let \( C \) and \( S \) represent the captain’s and sergeant’s
order, respectively. Then \( C \) and \( S \) can each take three values, 1, \(-1\), or 0, depending on the
order (attack, retreat, no order). The actual value of \( C \) and \( S \) is determined by the context.

\( P \) describes what the private does; as the story suggests, \( P = C \) if \( C \neq 0 \); otherwise \( P = S \). In
the actual context, \( C = S = P = 1 \).

\( C = 1 \) is a but-for cause of \( P = 1 \); setting \( C \) to \(-1\) would result in \( P = -1 \). Thus, \( C = 1 \)
is a cause of \( P = 1 \) according to all variants of the HP definition. \( S = 1 \) is not a cause of
\( B = 1 \) according to the modified HP definition; changing \( S \) to 0 while keeping \( C \) fixed at
1 will not affect the private’s action. However, \( S = 1 \) is a cause of \( P = 1 \) according to the
original and updated HP definition, with witness \((\{C\}, 0, 0)\) (i.e., in the witness world, \( C = 0 \)
and \( S = 0 \)): if the captain does nothing, then what the private does is completely determined
by the sergeant’s action.

Next, suppose that the captain and sergeant can only order an attack or a retreat; that is,
the range of each of \( C \) and \( S \) is \{\(-1, 1\)\}. Now it is easy to check that \( C = 1 \) is the only
cause of \( P = 1 \) according to all the variants of the HP definition. For the original and updated
definition, \( S = 1 \) is not a cause because there is no setting of \( C \) that will allow the sergeant’s
action to make a difference.

Finally, suppose that the captain and sergeant can only order an attack or do nothing; that is,
the range of each of \( C \) and \( S \) is \{0, 1\}. \( C = 1 \) and \( S = 1 \) are both causes of \( P = 1 \) according
to the original and updated HP definition, using the same argument used to show that \( S = 1 \)
was a cause when it was also possible to order a retreat. Neither is a cause according to the
modified HP definition, but \( C = 1 \land S = 1 \) is a cause, so both \( C = 1 \) and \( S = 1 \) are parts of
a cause.

Note that if the range of the variables \( C \) and \( S \) is \{0, 1\}, then there is no way to capture the
fact that in the case of conflicting orders, the private obeys the captain. In this setting, we can
capture the fact that the private is “really” obeying the captain if both the captain and sergeant
give orders by adding a new variable \( SE \) that captures the sergeant’s “effective” order. If the
captain does not issue any orders (i.e., if \( C = 0 \)), then \( SE = S \). If the captain does issue
an order, then \( SE = 0 \); the sergeant’s order is effectively blocked. In this model, \( P = C \) if
\( C \neq 0 \); otherwise, \( P = SE \). The causal network for this model is given in Figure 2.7.

In this model, the captain causes the private to advance, but the sergeant does not, according
to all variants of the HP definition. To see that the captain is a cause according to the modified
HP definition, hold \( SE \) fixed at 0 and set \( C = 0 \); clearly \( P = 0 \). By Theorem 2.2.3, \( C = 1 \) is
Chapter 2. The HP Definition of Causality

Figure 2.7: A model of commands that captures trumping.

A cause of $P = 1$ according to the original and updated definition as well. To show that $S = 1$ is not (part of) a cause according to all variants of the HP definition, it suffices to show it is not a cause according to the original HP definition. Suppose that we want to argue that $S = 1$ causes $P = 1$. The obvious thing to do is to take $\vec{W} = \{C\}$ and $\vec{Z} = \{S, SE, P\}$. However, this choice does not satisfy AC2(b°): if $C = 0$, $SE = 0$ (its original value), and $S = 1$, then $P = 0$, not 1. The key point is that this more refined model allows a setting where $C = 0$, $S = 1$, and $P = 0$ (because $SE = 0$). That is, despite the sergeant issuing an order to attack and the captain being silent, the private does nothing.

The final example in this section touches on issues of responsibility in the law.

**Example 2.3.8** Suppose that two companies both dump pollutant into the river. Company $A$ dumps 100 kilograms of pollutant; company $B$ dumps 60 kilograms. The fish in the river die. Biologists determine that $k$ kilograms of pollutant suffice for the fish to die. Which company is the cause of the fish dying if $k = 120$, if $k = 80$, and if $k = 50$?

It is easy to see that if $k = 120$, then both companies are causes of the fish dying, according to all three definitions (each company is a but-for cause of the outcome). If $k = 50$, then each company is still a cause according to the original and updated HP definition. For example, to see that company $B$ is a cause, we consider the contingency where company $A$ does not dump any pollutant. Then the fish die if company $B$ pollutes, but they survive if $B$ does not pollute. With the modified definition, neither company individually is a cause; there is no variable that we can hold at its actual value that would make company $A$ or company $B$ a but-for cause. However, both companies together are the cause.

The situation gets more interesting if $k = 80$. Now the modified definition says that only $A$ is a cause; if $A$ dumps 100 tons of pollutant, then what $B$ does has no impact. The original and updated definition also agree that $A$ is a cause if $k = 80$. Whether $B$ is a cause depends on the possible amounts of pollutant that $A$ can dump. If $A$ can dump only 0 or 100 kilograms of pollutant, then $B$ is not a cause; no setting of $A$’s action can result in $B$’s action making a difference. However, if $A$ can dump some amount between 20 and 79 kilograms, then $B$ is a cause.

It’s not clear what the “right” answer should be here if $k = 80$. The law typically wants to declare $B$ a contributing cause to the death of the fish (in addition to $A$), but should this depend on the amount of pollutant that $A$ can dump? As we shall see in Section 6.2, thinking in terms of responsibility and blame helps clarify the issue (see Example 6.2.5). Under minimal assumptions about how likely various amounts of pollutant are to be dumped, $B$ will get some degree of blame according to the modified definition, even when it is not a cause.
2.4 Transitivity

Example 2.4.1 Now consider the following modification of Example 2.3.5.

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 $M_B$ for this story is straightforward. There are three 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 feels fine both Tuesday morning and Wednesday morning; 1 if Billy feels sick Tuesday morning, fine Wednesday morning; 2 if Billy feels sick both Tuesday and Wednesday morning; 3 if Billy feels fine Tuesday morning and is 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; the causal network is shown in Figure 2.8.

![Figure 2.8: Billy’s medical condition.](http://direct.mit.edu/books/book/chapter-pdf/269281/9780262336611_cab.pdf)
Chapter 2. The HP Definition of Causality

HP definition. It suffices to show that it is not a cause according the original HP definition. This follows because setting $MT = 0$ cannot result in Billy dying, not matter how we fix $TT$.

This shows that causality is not transitive according to the HP definition. Although $MT = 1$ is a cause of $TT = 0$ and $TT = 0$ is a cause of $BMC = 0 \lor BMC = 1 \lor BMC = 2$, $MT = 1$ is not a cause of $BMC = 0 \lor BMC = 1 \lor BMC = 2$. Nor is causality closed under right weakening; that is, replacing a conclusion by something it implies. If $A$ is a cause of $B$ and $B$ logically implies $B'$, then $A$ may not be a cause of $B'$. In this case, $MT = 1$ is a cause of $BMC = 0$, which logically implies $BMC = 0 \lor BMC = 1 \lor BMC = 2$, which is not caused by $MT = 1$.

Although this example may seem somewhat forced, there are many quite realistic examples of lack of transitivity with exactly the same structure. Consider the body’s homeostatic system. An increase in external temperature causes a short-term increase in core body temperature, which in turn causes the homeostatic system to kick in and return the body to normal core body temperature shortly thereafter. But if we say that the increase in external temperature happened at time 0 and the return to normal core body temperature happened at time 1, we certainly would not want to say that the increase in external temperature at time 0 caused the body temperature to be normal at time 1!

There is another reason that causality is intransitive, which is illustrated by the following example.

**Example 2.4.2** Suppose that a dog bites Jim’s right hand. Jim was planning to detonate a bomb, which he normally would do by pressing the button with his right forefinger. Because of the dog bite, he presses the button with his left forefinger. The bomb still goes off.

Consider the causal model with variables $DB$ (the dog bites, with values 0 and 1), $P$ (the press of the button, with values 0, 1, and 2, depending on whether the button is not pressed at all, pressed with the right hand, or pressed with the left hand), and $B$ (the bomb goes off). We have the obvious equations: $DB$ is determined by the context, $P = DB + 1$, and $B = 1$ if $P$ is either 1 or 2. In the context where $DB = 1$, it is clear that $DB = 1$ is a but-for cause of $P = 2$ (if the dog had not bitten, $P$ would have been 1), and $P = 2$ is a but-for cause of $B = 1$ (if $P$ were 0, then $B$ would be 0), but $DB = 1$ is not a cause of $P = 1$. Regardless of whether the dog had bitten Jim, the button would have been pressed, and the bomb would have detonated.

In retrospect, the failure of right weakening is not so surprising. Taking true to be a tautology, if $A$ is a cause of $B$, then we do not want to say that $A$ is a cause of true, although $B$ logically implies true. However, the failure of transitivity is quite surprising. Indeed, despite Examples 2.4.1 and 2.4.2, it seems natural to think of causality as transitive. People often think in terms of causal chains: $A$ caused $B$, $B$ caused $C$, $C$ caused $D$, and therefore $A$ caused $D$, where transitivity seems natural (although the law does not treat causality as transitive in long causal chains; see Section 3.4.4). Not surprisingly, there are some definitions of causality that require causality to be transitive; see the notes at the end of the chapter for details.

Why do we feel that causality is transitive? I believe that this is because, in typical settings, causality is indeed transitive. I give below two simple sets of conditions that are sufficient to guarantee transitivity. Because I expect that these conditions apply in many cases, it may...
2.4 Transitivity

explain why we naturally generalize to thinking of causality as transitive and are surprised when it is not.

For the purposes of this section, I restrict attention to but-for causes. This is what the law has focused on and seems to be the situation that arises most often in practice (although I will admit I have only anecdotal evidence to support this); certainly the law has done reasonably well by considering only but-for causality. Restricting to but-for causality has the further advantage that all the variants of the HP definition agree on what counts as a cause. (Thus, in this section, I do not specify which variant of the definition I am considering.) Restricting to but-for causality does not solve the transitivity problem. As Examples 2.4.1 and 2.4.2 already show, even if \( X_1 = x_1 \) is a but-for cause of \( X_2 = x_2 \) and \( X_2 = x_2 \) is a but-for cause of \( X_3 = x_3 \), it may not be the case that \( X_1 = x_1 \) is a cause of \( X_3 = x_3 \).

The first set of conditions assumes that \( X_1, X_2, \) and \( X_3 \) each has a default setting. Such default settings make sense in many applications; “nothing happens” can be often taken as the default. Suppose, for example, a billiards expert hits ball \( A \), causing it to hit ball \( B \), causing it to carom into ball \( C \), which then drops into the pocket. In this case, we can take the default setting for the shot to be the expert doing nothing and the default setting for the balls to be that they are not in motion. Let the default setting be denoted by the value 0.

**Proposition 2.4.3** Suppose that

\[
(a) \quad X_1 = x_1 \text{ is a but-for cause of } X_2 = x_2 \text{ in } (M, \vec{u}),
\]

\[
(b) \quad X_2 = x_2 \text{ is a but-for cause of } X_3 = x_3 \text{ in } (M, \vec{u}),
\]

\[
(c) \quad x_3 \neq 0,
\]

\[
(d) \quad (M, \vec{u}) \models [X_1 \leftarrow 0](X_2 = 0), \text{ and}
\]

\[
(e) \quad (M, \vec{u}) \models [X_1 \leftarrow 0, X_2 \leftarrow 0](X_3 = 0).
\]

Then \( X_1 = x_1 \) is a but-for cause of \( X_3 = x_3 \) in \( (M, \vec{u}) \).

**Proof:** If \( X_2 = 0 \) in the unique solution to the equations in the causal model \( M_{X_1,\leftarrow 0} \) in context \( \vec{u} \) and \( X_3 = 0 \) in the unique solution to the equations in \( M_{X_1,\leftarrow 0, X_2,\leftarrow 0} \) in context \( \vec{u} \), then it is immediate that \( X_3 = 0 \) in the unique solution to the equations in \( M_{X_1,\leftarrow 0} \) in context \( \vec{u} \). That is, \( (M, \vec{u}) \models [X_1 \leftarrow 0](X_3 = 0) \). It follows from assumption (a) that \( (M, \vec{u}) \models X_1 = x_1 \). We must thus have \( x_1 \neq 0 \); otherwise, \( (M, \vec{u}) \models X_1 = 0 \land [X_1 \leftarrow 0](X_3 = 0) \), so \( (M, \vec{u}) \models X_3 = 0 \), which contradicts assumptions (b) and (c). Thus, \( X_1 = x_1 \) is a but-for cause of \( X_3 = x_3 \), since the value of \( X_3 \) depends counterfactually on that of \( X_1 \).

Although the conditions of Proposition 2.4.3 are clearly rather specialized, they arise often in practice. Conditions (d) and (e) say that if \( X_1 \) remains in its default state, then so will \( X_2 \), and if both \( X_1 \) and \( X_2 \) remain in their default states, then so will \( X_3 \). Put another way, this says that if the reason for \( X_2 \) not being in its default state is \( X_1 \) not being in its default state, and the reason for \( X_3 \) not being in its default state is \( X_1 \) and \( X_2 \) both not being in their default states. The billiard example can be viewed as a paradigmatic example of when these conditions apply. It seems reasonable to assume that if the expert does not shoot, then ball \( A \)}
does not move; and if the expert does not shoot and ball $A$ does not move (in the context of interest), then ball $B$ does not move, and so on.

Of course, the conditions on Proposition 2.4.3 do not apply in either Example 2.4.1 or Example 2.4.2. The obvious default values in Example 2.4.1 are $MT = TT = 0$, but the equations say that in all contexts $\vec{u}$ of the causal model $M_B$ for this example, we have $(M_B, \vec{u}) \models [MT \leftarrow 0](TT = 1)$. In the second example, if we take $DB = 0$ and $P = 0$ to be the default values of $DB$ and $P$, then in all contexts $\vec{u}$ of the causal model $M_D$, we have $(M_D, \vec{u}) \models [DB \leftarrow 0](P = 1)$.

While Proposition 2.4.3 is useful, there are many examples where there is no obvious default value. When considering the body’s homeostatic system, even if there is arguably a default value for core body temperature, what is the default value for the external temperature? But it turns out that the key ideas of the proof of Proposition 2.4.3 apply even if there is no default value. Suppose that $X_1 = x_1$ is a but-for cause of $X_2 = x_2$ in $(M, \vec{u})$ and $X_2 = x_2$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$. Then to get transitivity, it suffices to find values $x'_1$, $x'_2$, and $x'_3$ such that $x_3 \neq x'_3$, $(M, \vec{u}) \models [X_1 \leftarrow x'_1](X_2 = x'_2)$, and $(M, \vec{u}) \models [X_1 \leftarrow x'_1, X_2 \leftarrow x'_2](X_3 = x'_3)$. The argument in the proof of Proposition 2.4.3 (formalized in Lemma 2.10.2) shows that $(M, \vec{u}) \models [X_1 \leftarrow x'_1](X_3 = x'_3)$. It then follows that $X_1 = x_1$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$. In Proposition 2.4.3, $x'_1$, $x'_2$, and $x'_3$ were all 0, but there is nothing special about the fact that 0 is a default value here. As long as we can find some values $x'_1$, $x'_2$, and $x'_3$, these conditions apply. I formalize this as Proposition 2.4.4, which is a straightforward generalization of Proposition 2.4.3.

**Proposition 2.4.4** Suppose that there exist values $x'_1$, $x'_2$, and $x'_3$ such that

(a) $X_1 = x_1$ is a but-for cause of $X_2 = x_2$ in $(M, \vec{u})$,

(b) $X_2 = x_2$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$,

(c) $x_3 \neq x'_3$,

(d) $(M, \vec{u}) \models [X_1 \leftarrow x'_1](X_2 = x'_2)$ (i.e., $(X_2)_{x'_1}(\vec{u}) = x'_2$), and

(e) $(M, \vec{u}) \models [X_1 \leftarrow x'_1, X_2 \leftarrow x'_2](X_3 = x'_3)$ (i.e., $(X_3)_{x'_1,x'_2}(\vec{u}) = x'_3$).

Then $X_1 = x_1$ is a but-for cause of $X_3 = x_3$ in $(M, \vec{u})$.

To see how these ideas apply, suppose that a student receives an A+ in a course, which causes her to be accepted at Cornell University (her top choice, of course!), which in turn causes her to move to Ithaca. Further suppose that if she had received an A in the course, she would have gone to university $U_1$ and as a result moved to city $C_1$, and if she had gotten anything else, she would have gone to university at $U_2$ and moved to city $C_2$. This story can be captured by a causal model with three variables: $G$ for her grade, $U$ for the university she goes to, and $C$ for the city she moves to. There are no obvious default values for any of these three variables. Nevertheless, we have transitivity here: The student’s A+ was a cause of her being accepted at Cornell, and being accepted at Cornell was a cause of her move to Ithaca; it seems like a reasonable conclusion that the student’s A+ was a cause of her move to Ithaca. Indeed, transitivity follows from Proposition 2.4.4. We can take the student getting an A to be
2.4 Transitivity

Because $x'_1$, the student being accepted at university $U_1$ to be $x'_2$, and the student moving to $C_1$ to be $x'_3$ (assuming that $U_1$ is not Cornell and that $C_1$ is not Ithaca, of course).

The conditions provided in Proposition 2.4.4 are not only sufficient for causality to be transitive, they are necessary as well, as the following result shows.

**Proposition 2.4.5** If $X_1 = x_1$ is a but-for cause of $X_3 = x_3$ in $(M, \bar{u})$, then there exist values $x'_1$, $x'_2$, and $x'_3$ such that $x_3 \neq x'_3$, $(M, \bar{u}) \models [X_1 \leftarrow x'_1](X_2 = x'_2)$, and $(M, \bar{u}) \models [X_1 \leftarrow x'_1, X_2 \leftarrow x'_2](X_3 = x'_3)$.

**Proof:** Since $X_1 = x_1$ is a but-for cause of $X_3 = x_3$ in $(M, \bar{u})$, there must exist a value $x'_1 \neq x_1$ such that $(M, \bar{u}) \models [X_1 \leftarrow x'_1](X_3 \neq x_3)$. Let $x'_2$ and $x'_3 \neq x_3$ be such that $(M, \bar{u}) \models [X_1 \leftarrow x'_1](X_2 = x'_2 \land X_3 = x'_3)$. It easily follows that $(M, \bar{u}) \models [X_1 \leftarrow x'_1, X_2 \leftarrow x'_2](X_3 = x'_3)$.

In light of Propositions 2.4.4 and 2.4.5, understanding why causality is so often taken to be transitive comes down to finding sufficient conditions to guarantee the assumptions of Proposition 2.4.4. I now present a set of conditions sufficient to guarantee the assumptions of Proposition 2.4.4, motivated by the two examples showing that causality is not transitive. To deal with the problem in Example 2.4.2, I require that for every value $x'_2$ in the range of $X_2$, there is a value $x'_1$ in the range of $X_1$ such that $(M, \bar{u}) \models [X_1 \leftarrow x'_1](X_2 = x'_2)$. This requirement holds in many cases of interest; it is guaranteed to hold if $X_1 = x_1$ is a but-for cause of $X_2 = x_2$ and $X_2$ is binary (since but-for causality requires that two different values of $X_1$ result in different values of $X_2$). But this requirement does not hold in Example 2.4.2; no setting of $DB$ can force $P$ to be 0.

Imposing this requirement still does not deal with the problem in Example 2.4.1. To do that, we need only one more condition: $X_2$ must lie on every causal path from $X_1$ to $X_3$. Roughly speaking, this says that all the influence of $X_1$ on $X_3$ goes through $X_2$. This condition does not hold in Example 2.4.1: as Figure 2.8 shows, there is a direct causal path from $MT$ to $BMC$ that does not include $TT$. However, this condition does hold in many cases of interest. Going back to the example of the student’s grade, the only way that the student’s grade can influence which city the student moves to is via the university that accepts the student.

To make this precise, I first need to define causal path. A causal path in a causal setting $(M, \bar{u})$ is a sequence $(Y_1, \ldots, Y_k)$ of variables such that $Y_{j+1}$ depends on $Y_j$ in context $\bar{u}$ for $j = 1, \ldots, k - 1$. Since there is an edge between $Y_j$ and $Y_{j+1}$ in the causal network for $M$ (assuming that we fix context $\bar{u}$) exactly if $Y_{j+1}$ depends on $Y_j$, a causal path is just a path in the causal network. A causal path in $(M, \bar{u})$ from $X_1$ to $X_2$ is just a causal path whose first node is $X_1$ and whose last node is $X_2$. Finally, $Y$ lies on a causal path in $(M, \bar{u})$ from $X_1$ to $X_2$ if $Y$ is a node (possibly $X_1$ or $X_2$) on a causal path in $(M, \bar{u})$ from $X_1$ to $X_2$.

The following result summarizes the second set of conditions sufficient for transitivity. (Recall that $\mathcal{R}(X)$ denotes the range of the variable $X$.)

**Proposition 2.4.6** Suppose that $X_1 = x_1$ is a but-for cause of $X_2 = x_2$ in the causal setting $(M, \bar{u}), X_2 = x_2$ is a but-for cause of $X_3 = x_3$ in $(M, \bar{u})$, and the following two conditions hold:

(a) for every value $x'_2 \in \mathcal{R}(X_2)$, there exists a value $x'_1 \in \mathcal{R}(X_1)$ such that $(M, \bar{u}) \models [X_1 \leftarrow x'_1](X_2 = x'_2)$ (i.e., $(X_2)_{x'_1(\bar{u})} = x'_2$);
(b) $X_2$ is on every causal path in $(M, \bar{u})$ from $X_1$ to $X_3$.

Then $X_1 = x_1$ is a but-for cause of $X_3 = x_3$.

The proof of Proposition 2.4.6 is not hard, although we must be careful to get all the details right. The high-level idea of the proof is easy to explain. Suppose that $X_2 = x_2$ is a but-for cause of $X_3 = x_3$ in $(M, \bar{u})$. Then there must be some values $x_2 \neq x_2'$ and $x_3 \neq x_3'$ such that $(M, \bar{u}) \models [X_2 \leftarrow x_2'](X_3 = x_3')$. By assumption, there exists a value $x_1' \in \mathcal{R}(X_1)$ such that $(M, \bar{u}) \models [X_1 \leftarrow x_1'](X_2 = x_2')$. The requirement that $X_2$ is on every causal path from $X_1$ to $X_3$ guarantees that $[X_2 \leftarrow x_2'](X_3 = x_3)$ implies $[X_1 \leftarrow x_1', X_2 \leftarrow x_2'](X_3 = x_3)$ in $(M, \bar{u})$ (i.e., $(X_3)_{x_2'}(\bar{u}) = x_3$ implies $(X_3)_{x_1', x_2'}(\bar{u}) = x_3$). Roughly speaking, $X_2$ “screens off” the effect of $X_1$ on $X_3$, since it is on every causal path from $X_1$ to $X_3$. Now we can apply Proposition 2.4.4. I defer the formal argument to Section 2.10.2.

It is easy to construct examples showing that the conditions of Proposition 2.4.6 are not necessary for causality to be transitive. Suppose that $X_1 = x_1$ causes $X_2 = x_2$, $X_2 = x_2$ causes $X_3 = x_3$, and there are several causal paths from $X_1$ to $X_3$. Roughly speaking, the reason that $X_1 = x_1$ may not be a but-for cause of $X_3 = x_3$ is that the effects of $X_1$ on $X_3$ may “cancel out” along the various causal paths. This is what happens in the homeostasis example. If $X_2$ is on all the causal paths from $X_1$ to $X_3$, then as we have seen, all the effect of $X_1$ on $X_3$ is mediated by $X_2$, so the effect of $X_1$ on $X_3$ on different causal paths cannot “cancel out”. But even if $X_2$ is not on all the causal paths from $X_1$ to $X_3$, the effects of $X_1$ on $X_3$ may not cancel out along the causal paths, and $X_1 = x_1$ may still be a cause of $X_3 = x_3$. That said, it seems difficult to find a weakening of the condition in Proposition 2.4.6 that is simple to state and suffices for causality to be transitive.

### 2.5 Probability and Causality

In general, an agent trying to determine whether $A$ is a cause of $B$ may not know the exact model or context. An agent may be uncertain about whether Suzy and Billy both threw, or just Suzy threw; he may be uncertain about who will hit first if both Suzy and Billy throw (or perhaps they hit simultaneously). In a perhaps more interesting setting, at some point, people were uncertain about whether smoking caused cancer or whether smoking and cancer were both the outcome of the same genetic problem, which would lead to smoking and cancer being correlated, but no causal relationship between the two. To deal with such uncertainty, an agent can put a probability on causal settings. For each causal setting $(M, \bar{u})$, the agent can determine whether $A$ is a cause of $B$ in $(M, \bar{u})$, and so compute the probability that $A$ is a cause of $B$. Having probabilities on causal settings will be relevant in the discussion of blame in Chapter 6. Here I focus on what seems to be a different source of uncertainty: uncertainty in the equations. In causal models, all the equations are assumed to be deterministic. However, as was observed, in many cases, it seems more reasonable to think of outcomes as probabilistic. So rather than thinking of Suzy’s rock as definitely hitting the bottle if she throws, we can think of her as hitting with probability .9. That is, rather than taking Suzy to always be accurate, we can take her to be accurate with some probability.

The first step in considering how to define causality in the presence of uncertain outcomes is to show how this uncertainty can be captured in the formal framework. Earlier, I assumed
that, for each endogenous variable \( X \), there was a deterministic function \( F_X \) that described the value of \( X \) as a function of all the other variables. Now, rather than assuming that \( F_X \) returns a particular value, I assume that it returns a distribution over the values of \( X \) given the values of all other variables. This is perhaps easiest to understand by means of an example. In the rock-throwing example, rather than assuming that if Suzy throws a rock, she definitely hits the bottle, suppose we assume that she only only hits it with probability .9 and misses with probability .1. Suppose that if Billy throws and Suzy hasn’t hit the bottle, then Billy will hit it with probability .8 and miss with probability .2. Finally, for simplicity, assume that, if hit by either Billy or Suzy, the bottle will definitely shatter.

The next two paragraphs show how the probabilistic version of this story can be modeled formally in the structural-equations framework and can be skipped on a first reading. Recall that for the more sophisticated model \( M'_{RT} \) of the (deterministic) rock-throwing story, although I earlier wrote the equation for \( SH \) as \( SH = ST \), this is really an abbreviation for the function \( F_{SH}(i_1, i_2, i_3, i_4, i_5) = i_2; \) \( F_{SH} \) is a function of the values of the exogenous variable \( U \) (which is not described in the causal network but determines the values of \( BT \) and \( ST \)) and the endogenous variables \( ST, BT, BH, \) and \( BS \). These values are given by \( i_1, \ldots, i_5 \). The equation \( SH = ST \) says that the value of \( SH \) depends only on the value of \( ST \) and is equal to it; that explains the output \( i_2 \). For the probabilistic version of the story, \( F_{SH}(i_1, 1, i_3, i_4, i_5) \) is the probability distribution that places probability .9 on the value 1 and probability .1 on the value 0. In words, this says that if Suzy throws (\( ST = 1 \)), then no matter what the values \( i_1, i_3, i_4, \) and \( i_5 \) of \( U, BT, BH, \) and \( BS \) are, respectively, \( F_{SH}(i_1, 1, i_3, i_4, i_5) \) is the probability distribution that puts probability .9 on the event \( SH = 1 \) and probability .1 on the event \( SH = 0 \). This captures the fact that if Suzy throws, then she has a probability .9 of hitting the bottle (no matter what Billy does). Similarly, \( F_{SH}(i_1, 0, i_3, i_4, i_5) \) is the distribution that places probability 1 on the value 0: if Suzy doesn’t throw, then she certainly won’t hit the bottle, so the event \( SH = 0 \) has probability 1.

Similarly, in the deterministic model of the rock-throwing example, The function \( F_{BH} \) is such that \( F_{BH}(i_1, i_2, i_3, i_4, i_5) = 1 \) if \( (i_3, i_4) = (1, 0) \), and is 0 otherwise. That is, taking \( i_1, \ldots, i_5 \) to be the values of \( U, ST, BT, SH, \) and \( BS \), respectively, Billy’s throw hits the bottle only if he throws \( i_3 = 1 \) and Suzy doesn’t hit the bottle \( i_4 = 0 \). For the probabilistic version of the story, \( F_{BH}(i_1, i_2, i_3, i_4, i_5) \) is the distribution that puts probability .8 on 1 and probability .2 on 0; for all other values of \( i_3 \) and \( i_4 \), \( F_{BH}(i_1, i_2, 1, 0, i_5) \) puts probability 1 on 0. (Henceforth, I just describe the probabilities in words rather than using the formal notation.)

The interpretation of these probabilities is similar to the interpretation of the deterministic equations we have used up to now. For example, the fact that Suzy’s rock hits with probability .9 does not mean that the probability of Suzy’s rock hitting conditional on her throwing is .9; rather, it means that if there is an intervention that results in Suzy throwing, the probability of her hitting is .9. The probability of rain conditional on a low barometer reading is high. However, intervening on the barometer reading, say, by setting the needle to point to a low reading, does not affect the probability of rain.

There have been many attempts to give a definition of causality that takes probability into account. They typically take \( A \) to be a cause of \( B \) if \( A \) raises the probability of \( B \). I am going to take a different approach here. I take a standard technique in computer science to “pull
out the probability”, allowing me to convert a single causal setting where the equations are probabilistic to a probability over causal settings, where in each causal setting, the equations are deterministic. This, in turn, will allow me to avoid giving a separate definition of probabilistic causality. Rather, I will be able to use the definition of causality already given for deterministic models and talk about the probability of causality, that is, the probability that $A$ is a cause of $B$. As we shall see, this approach seems to deal naturally with a number of problems regarding probabilistic causality that have been raised in the literature.

The assumption is that the equations would be deterministic if we knew all the relevant details. This assumption is probably false at the quantum level; most physicists believe that at this level, the world is genuinely probabilistic. But for the macroscopic events that I focus on in this book (and that most people are interested in when applying causality judgments), it seems reasonable to view the world as fundamentally deterministic. With this viewpoint, if Suzy hits the rock with probability .9, then there must be (perhaps poorly understood) reasons that cause her to miss: an unexpected gust of wind or a momentary distraction throwing off her aim. We can “package up” all these factors and make them exogenous. That is, we can have an exogenous variable $U'$ with values either 0 or 1 depending on whether the features that cause Suzy to miss are present, where $U' = 0$ with probability .1 and $U' = 1$ with probability .9. By introducing such a variable $U'$ (and a corresponding variable for Billy), we have essentially pulled the probability out of the equations and put it into the exogenous variable.

I now show in more detail how this approach plays out in the context of the rock-throwing example with the probabilities given above. Consider the causal setting where Suzy and Billy both throw. We may be interested in knowing the probability that Suzy or Billy will be a cause of the bottle shattering if we don’t know whether the bottle will in fact shatter, or in knowing the probability that Suzy or Billy is the cause if we know that the bottle actually did shatter. (Notice that in the former case, we are asking about an effect of a (possible) cause, whereas in the latter case, we are asking about the cause of an effect.) When we pull out the probability, there are four causal models that arise here, depending on whether Suzy’s rock hits the bottle and whether Billy’s rock hits the bottle if Suzy’s rock does not hit. Specifically, consider the following four models, in a context where both Suzy and Billy throw rocks:

- $M_1$: Suzy’s rock hits and Billy’s rock would hit if Suzy missed;
- $M_2$: Suzy’s rock hits but Billy’s rock would not hit if Suzy missed;
- $M_3$: Suzy’s rock misses and Billy’s rock hits;
- $M_4$: neither rock hits.

Note that these models differ in their structural equations, although they involve the same exogenous and endogenous variables. $M_1$ is just the model considered earlier, described in Figure 2.3. Suzy’s throw is a cause in $M_1$ and $M_2$; Billy’s throw is a cause in $M_3$.

If we know that the bottle shattered, Suzy’s rock hit, and Billy’s didn’t, then the probability that Suzy is the cause is 1. We are conditioning on the model being either $M_1$ or $M_2$ because these are the models where Suzy’s rock hits, and Suzy is the cause of the bottle shattering in both. But suppose that all we know is that the bottle shattered, and we are trying to determine
the probability that Suzy is the cause. Now we can condition on the model being either $M_1$, $M_2$, or $M_3$, so if the prior probability of $M_4$ is $p$, then the probability that Suzy is a cause of the bottle shattering is $\frac{.9}{(1-p)}$: Suzy is a cause of the bottle shattering as long as she hits the bottle and it shatters (i.e., in models $M_1$ and $M_2$). Unfortunately, the story does not give us that the probability of $M_4$. It is $1-q$, where $q$ is the prior probability of $M_3$, but the story does not tell us that either. We can say more if we make a reasonable additional assumption: that whether Suzy’s rock hits is independent of whether Billy’s rock would hit if Suzy missed. Then $M_1$ has probability $\frac{.72}{.9} = \frac{.8}{.98} = 45/49$. Although it seems reasonable to view the outcomes “Suzy’s rock hits the bottle” and “Billy’s rock would have hit the bottle if Suzy’s hadn’t” as independent, they certainly do not have to be. We could, for example, assume that Billy and Suzy are perfectly accurate when there are no wind gusts and miss when there are wind gusts, and that wind gusts occur with probability $.7$. Then it would be the case that Suzy’s rock hits with probability $.7$ and that Billy’s rock would have hit with probability $.7$ if Suzy’s had missed, but there would have been only two causal models: the one that we considered in the deterministic case, which would be the actual model with probability $.7$, and the one where both miss. (A better model of this situation might have an exogenous variable $U’$ such that $U’ = 1$ if there are wind gusts and $U’ = 0$ if there aren’t, and then have deterministic equations for $SH$ and $BH$.) The key message here is that, in general, further information might be needed to determine the probability of causality beyond what is provided by probabilistic structural equations as defined above.

A few more examples should help illustrate how this approach works and how much we can say.

**Example 2.5.1** Suppose that a doctor treats Billy on Monday. The treatment will result in Billy recovering on Tuesday with probability $.9$. But Billy might also recover with no treatment due to some other factors, with independent probability $.1$. By “independent probability” here, I mean that conditional on Billy being treated on Monday, the events “Billy recovers due to the treatment” and “Billy recovers due to other factors” are independent; thus, the probability that Billy will not recover on Tuesday given that he is treated on Monday is $(.1)(.9)$. Now, in fact, Billy is treated on Monday and he recovers on Tuesday. Is the treatment the cause of Billy getting better?

The standard answer is yes, because the treatment greatly increases the likelihood of Billy getting better. But there is still the nagging concern that Billy’s recovery was not due to the treatment. He might have gotten better if he had just been left alone.

Formally, we have a causal model with three endogenous binary variables: $MT$ (the doctor treats Billy on Monday), $OF$ (the other factors that would make Billy get better occur), and $BMC$ (Billy’s medical condition, which we can take to be 1 if he gets better and 0 otherwise). The exogenous variable(s) determine $MT$ and $OF$. The probability that $MT = 1$ does not matter for this analysis; for definiteness, assume that $MT = 1$ with probability $.8$. According to the story, $OF = 1$ with probability $.9$. Finally, $BMC = 1$ if $OF = 1$, $BMC = 0$ if $MT = OF = 0$, and $BMC = 1$ with probability $.9$ if $MT = 1$ and $OF = 0$. Again, after we pull out the probability, there are eight models, which differ in whether $MT$ is set to 0 or
Chapter 2. The HP Definition of Causality

1, whether $OF$ is set to 0 or 1, and whether $BMC = 1$ when $MT = 1$ and $OF = 0$. Treating these choices as independent (as is suggested by the story), it is straightforward to calculate the probability of each of the eight models. Since we know that Billy actually did get better and was treated by his doctor, we ignore the four models where $MT = 0$ and the model where $MT = 1$, $OF = 0$, and $BMC = 0$ if $MT = 1$ and $OF = 0$. The remaining three models have probability $0.8(1 - (0.9)(0.1))$. $MT = 1$ is a cause of $BMC = 0$ in two of the three models; it is not a cause only in the model where $OF = 1$, $MT = 1$, but $BMC$ would be 0 if $OF$ were set to 0. This latter model has prior probability $0.8(0.1)(0.1)$. Straightforward calculations now show that the treatment cause of Billy’s recovery with probability $\frac{9}{1 - (0.9)(0.1)} = 90/91$. Similarly, the probability that Billy’s recovery was caused by other factors is $\frac{1}{1 - (0.1)(0.9)} = 10/91$. These probabilities sum to more than 1 because with probability $0.9(0.1)(0.9) = 9/91$, both are causes; Billy’s recovery is overdetermined.

Now suppose that we can perform a (somewhat expensive) test to determine whether the doctor’s treatment actually caused Billy to get better. Again, the doctor treats Billy and Billy recovers, but the test is carried out, and it is shown that the treatment was not the cause. It is still the case that the treatment significantly raises the probability of Billy recovering, but now we certainly don’t want to call the treatment the cause. And, indeed, it is not. Conditioning on the information, we can discard three of the four models. The only one left is the one in which Billy’s recovery is due to other factors (with probability 1). Moreover, this conclusion does not depend on the independence assumption. □

Example 2.5.2 Now suppose that there is another doctor who can treat Billy on Monday, with different medication. Both treatments are individually effective with probability 0.9. Unfortunately, if Billy gets both treatments and both are effective, then, with probability 0.8, there will be a bad interaction, and Billy will die. For this version of the story, suppose that there are no other factors involved; if Billy gets neither treatment, he will not recover. Further suppose that we can perform a test to see whether a treatment was effective. Suppose that in fact both doctors treat Billy and, despite that, Billy recovers on Tuesday. Without further information, there are three relevant models: in the first, only the first doctor’s medication was effective; in the second, only the second doctor’s medication was effective; in the third, both are effective. If further testing shows that both treatments were effective, then there is no uncertainty; we are down to just one model: the third one. According to the original and updated HP definition, both treatments are the cause of Billy’s recovery, with probability 1. (According to the modified HP definition, they are both parts of the cause.) This is true despite the fact that, given that the first doctor treated Billy, the second doctor treating Billy significantly lowers the probability of Billy recovering (and symmetrically with the roles of the doctors reversed). □

Example 2.5.3 Now suppose that a doctor has treated 1,000 patients. They each would have had probability 0.9 of recovering even without the treatment. With the treatment, they recover with independent probability 0.1. In fact, 908 patients recovered, and Billy is one of them. What is the probability that the treatment was the cause of Billy’s recovery? Calculations similar to those in Example 2.5.1 show that it is 10/91. Standard probabilistic calculations show that there is quite a high probability that the treatment is a cause of at least one patient’s recovery. Indeed, there is quite a high probability that the treatment is the unique cause of at least one patient’s recovery. But we do not know which patient that is. □
Example 2.5.4  Again, Billy and Suzy throw rocks at a bottle. If either Billy or Suzy throws a rock, it will hit the bottle. But now the bottle is heavy and will, in general, not topple over if it is hit. If only Billy hits the bottle, it will topple over with probability .2; similarly, if only Suzy hits the bottle, it will topple over with probability .2. If they hit the bottle simultaneously, it will topple with probability .7. In fact, both Billy and Suzy throw rocks at the bottle, and it topples over. We are interested in the extent to which Suzy and Billy are the causes of the bottle toppling over.

For simplicity, I restrict attention to contexts where both Billy and Suzy throw, both hit, and the bottle topples over. After pulling out the probability, there are four causal models of interest, depending on what would have happened if just Suzy had hit the bottle and if just Billy had hit the bottle:

- $M_1$: the bottle would have toppled over if either only Suzy’s rock had hit it or if only Billy’s rock had hit it;
- $M_2$: the bottle would have toppled over if only Suzy’s rock had hit it but not if only Billy’s rock had hit it;
- $M_3$: the bottle would have toppled over if only Billy’s rock had hit it but not if only Suzy’s rock had hit it;
- $M_4$: the bottle would not have toppled over if either only Billy’s rock or only Suzy’s rock had hit it.

I further simplify by treating the outcomes “the bottle would have toppled over had only Suzy’s rock hit it” and “the bottle would have toppled over had only Billy’s rock hit it” as independent (even though neither may be independent of the outcome “the bottle would have toppled over had both Billy and Suzy’s rocks hit it”). With this assumption, the probability of $M_1$ (conditional on the bottle toppling over if both Suzy’s and Billy’s rock hit it) is .04, the conditional probability of $M_2$ is .16, the conditional probability of $M_3$ is .16, and the conditional probability of $M_4$ is .64.

It is easy to check that Suzy’s throw is a cause of the bottle toppling in models $M_1$, $M_2$, and $M_4$, but not in $M_3$. Similarly, Billy’s throw is a cause in models $M_1$, $M_3$, and $M_4$. They are both causes in models $M_1$ and $M_4$ according to the original and updated HP definition, as well as in $M_4$ according to the modified HP definition; in $M_1$, the conjunction $ST = 1 \land BT = 1$ is a cause according to the modified HP definition. Thus, the probability that Suzy’s throw is part of a cause of the bottle shattering is .84, and likewise for Billy. The probability that both are parts of causes is .68 (since this is the case in models $M_1$ and $M_4$).

As these examples show, the HP definition lets us make perfectly sensible statements about the probability of causality. There is nothing special about these examples; all other examples can be handled equally well. Perhaps the key message here is that there is no need to work hard on getting a definition of probabilistic causality, at least at the macroscopic level; it suffices to get a good definition of deterministic causality. But what about if we want to treat events at the quantum level, which seems inherently probabilistic? A case can still be made that it is psychologically useful to think deterministically. Whether the HP definitions can be extended successfully to an inherently probabilistic framework still remains open.
Even if we ignore issues at the microscopic level, there is still an important question to be addressed: In what sense is the single model with probabilistic equations equivalent to the set of deterministic causal models with a probability on them? In a naive sense, the answer is no. The examples above show that the probabilistic equations do not in general suffice to determine the probability of the deterministic causal models. Thus, there is a sense in which deterministic causal models with a probability on them carry more information than a single probabilistic model. Moreover, this extra information is useful. As we have seen, it may be necessary to determine the answer to questions such as “How probable is it that Billy is the cause of the bottle shattering?” when all we see is a shattered bottle and know that Billy and Suzy both threw. Such a determination can be quite important in, for example, legal cases. Although bottle shattering may not be of great legal interest, we may well want to know the probability of Bob being the cause of Charlie dying in a case where there are multiple potential causes.

If we make reasonable independence assumptions, then a probabilistic model determines a probability on deterministic models. Moreover, under the obvious assumptions regarding the semantics of causal models with probabilistic equations, formulas in the language described in Section 2.2.1 have the same probability of being true under both approaches. Consider Example 2.5.4 again. What is the probability of a formula like $[ST \leftarrow 0](BS = 1)$ (if Suzy hadn’t thrown, the bottle would have shattered)? Clearly, of the four deterministic models described in the discussion of this example, the formula is true only in $M_1$ and $M_3$; thus, it has probability 0.2. Although I have not given a semantics for formulas in this language in probabilistic causal models (i.e., models where the equations are probabilistic), if we assume that the outcomes “the bottle would have toppled over had only Suzy’s rock hit it” and “the bottle would have toppled over had only Billy’s rock hit it” are independent, as I assumed when assigning probabilities to the deterministic models, we would expect this statement to also be true in the probabilistic model. The converse holds as well: if a statement about the probability of causality holds in the probabilistic model, then it holds in the corresponding deterministic model.

This means that (under the independence assumption) as far as the causal language of Section 2.2.1 is concerned, the two approaches are equivalent. Put another way, to distinguish the two approaches, we need a richer language. More generally, although we can debate whether pulling out the probability “really” gives an equivalent formulation of the situation, to the extent that the notion of (probabilistic) causality is expressible in the language of Section 2.2.1, the two approaches are equivalent. In this sense, it is safe to reduce to deterministic models.

This observation leads to two further points. Suppose that we do not want to make the relevant independence assumptions. Is there a natural way to augment probabilistic causal models to represent this information (beyond just going directly to a probability over deterministic causal models)? See Section 5.1 for further discussion of this point.

Of course, we do not have to use probability to represent uncertainty; other representations of uncertainty could be used as well. Indeed, there is a benefit to using a representation that involves sets of probabilities. Since, in general, probabilistic structural equations do not determine the probability of the deterministic models that arise when we pull out the probability, all we have is a set of possible probabilities on the deterministic models. Given that, it seems reasonable to start with sets of probabilities on the equations as well. This
makes sense if, for example, we do not know the exact probability that Suzy’s rock would hit the bottle, but we know that it would hit it with probability between .6 and .8. It seems that the same general approach of pulling out the probabilities works if we represent uncertainty using sets of probabilities, but I have not explored this approach in detail.

2.6 Sufficient Causality

Although counterfactual dependence is a key feature of causality, people’s causality judgments are clearly influenced by another quite different feature: how sensitive the causality ascription is to changes in various other factors. If we assume that Suzy is very accurate, then her throwing a rock is a robust cause of the bottle shattering. The bottle will shatter regardless of whether Billy throws, even if the bottle is in a slightly different position and even if it is slightly windy. On the other hand, suppose instead that we consider a long causal chain like the following: Suzy throws a rock at a lock, causing it to open, causing the lion that was in the locked cage to escape, frightening the cat, which leapt up on to the table and knocked over the bottle, which then shattered. Although Suzy’s throw is still a but-for cause of the bottle shattering, the shattering is clearly sensitive to many other factors. If Suzy’s throw had not broken the lock, if the lion had run in a different direction, or if the cat had not jumped on the table, then the bottle would not have shattered. People seem inclined to assign less blame to Suzy’s throw in this case. I return to the issue of long causal chains in Sections 3.4.4 and provide a solution to it using the notion of blame in Section 6.2. Here I define a notion of sufficient causality that captures some of the intuitions behind insensitive causation. Sufficient causality also turns out to be related to the notion of explanation that I define in Chapter 7, so it is worth considering in its own right, although it will not be a focus of this book.

The key intuition behind the definition of sufficient causality is that not only does \( \vec{X} = \vec{x} \) suffice to bring about \( \varphi \) in the actual context (which is the intuition that AC2(b^o) and AC2(b^u) are trying to capture), but it also brings it about in other “nearby” contexts. Since the framework does not provide a metric on contexts, there is no obvious way to define nearby context. Thus, in the formal definition below, I start by considering all contexts. Since conjunction plays a somewhat different role in this definition than it does in the definition of actual causality (particularly in the modified HP definition), I take seriously here the intuition that what we really care about are parts of causes, rather than causes.

Again, there are three variants of the notion of sufficient causality, depending on which version of AC2 we use. The differences between these variants do not play a role in our discussion, so I just write AC2 below. Of course, if we want to distinguish, AC2 below can be replaced by either AC2(a) and AC2(b^o), AC2(a) and AC2(b^u), or AC2(a^m), depending on which version of the HP definition is being considered.

**Definition 2.6.1** \( \vec{X} = \vec{x} \) is a sufficient cause of \( \varphi \) in the causal setting \( (M, \vec{u}) \) if the following conditions hold:

**SC1.** \( (M, \vec{u}) \models (\vec{X} = \vec{x}) \) and \( (M, \vec{u}) \models \varphi \).

**SC2.** Some conjunct of \( \vec{X} = \vec{x} \) is part of a cause of \( \varphi \) in \( (M, \vec{u}) \). More precisely, there exists a conjunct \( X = x \) of \( \vec{X} = \vec{x} \) and another (possibly empty) conjunction \( \vec{Y} = \vec{y} \) such...
Chapter 2. The HP Definition of Causality

that $X = x \land \bar{Y} = \bar{y}$ is a cause of $\varphi$ in $(M, \bar{u})$; that is, AC1, AC2, and AC3 hold for $X = x \land \bar{Y} = \bar{y}$.

SC3. $(M, \bar{u}') \models [\bar{X} \leftarrow \bar{x}] \varphi$ for all contexts $\bar{u}'$.

SC4. $\bar{X}$ is minimal; there is no strict subset $\bar{X}'$ of $\bar{X}$ such that $\bar{X}' = \bar{x}'$ satisfies conditions SC1, SC2, and SC3, where $\bar{x}'$ is the restriction of $\bar{x}$ to the variables in $\bar{X}$.

SC3 is the key condition here; it says that $\bar{X} = \bar{x}$ suffices to bring about $\varphi$ in all contexts. Thus, in the case of the 11–0 victory, any set of six voters becomes a sufficient cause (no matter which variant of the HP definition we use), assuming that there are contexts corresponding to all possible voting configurations. This suggests that sufficient causality is related to the modified HP definition, which, in this case, would also take any subset of six voters to be a cause. However, this is misleading, as the next example shows.

Consider the forest-fire example again. In the disjunctive model, each of $L = 1$ and $MD = 1$ is a sufficient cause; in the conjunctive model, $L = 1 \land MD = 1$ is a sufficient cause, assuming that there is a context where there is no lightning and another where the arsonist does not drop a match. This is the case for all variants of the HP definition. Recall that this is just the opposite of what the modified HP definition does with actual causality; it would declare $L = 1 \land MD = 1$ the cause in the disjunctive model and both $L = 1$ and $MD = 1$ individually causes in the conjunctive model.

I wrote SC2 as I did so as to take seriously the notion that all we really care about when using the modified (and updated) definition is parts of causes. Thus, in the disjunctive model for the forest fire example, $L = 1$ and $MD = 1$ are both sufficient causes of the forest fire in the context $(1, 1)$ where there is both a lightning strike and a dropped match, even if we use AC2($a^m$) in SC2. By assumption, SC3 holds: both $[L \leftarrow 1](FF = 1)$ and $[MD \leftarrow 1](FF = 1)$ hold in all contexts. Moreover, both $L = 1$ and $MD = 1$ are parts of a cause of $FF = 1$ (namely, $L = 1 \land MD = 1$), so SC2 holds.

The forest-fire example shows that sufficient causality lets us distinguish what can be called joint causes from independent causes. In the disjunctive forest-fire model, the lightning and the dropped match can each be viewed as independent causes of the fire; each suffices to bring it about. In the conjunctive model, the lightning and the dropped match are joint causes; their joint action is needed to bring about the forest fire. The distinction between joint and independent causality seems to be one that people are quite sensitive to. Not surprisingly, it plays a role in legal judgments.

Going on with examples, consider the sophisticated rock-throwing model $M'_{RT}$ from Example 2.3.3 again. Assume that Suzy is always accurate; that is, she is accurate in all contexts where she throws, and would be accurate even in contexts where she doesn’t throw if (counterfactually) she were to throw; that is, assume that $[ST \leftarrow 1](BS = 1)$ holds in all contexts, even if Suzy does not actually throw. Then Suzy’s throw is a sufficient cause of the bottle shattering in the context that we have been considering, where both Billy and Suzy throw, and Suzy’s throw actually hits the bottle. Billy’s throw is not a sufficient cause in this context because it is not even (part of) a cause. However, if we consider the model $M''_{RT}$, which has a larger set of contexts, including ones where Suzy might throw and miss, then $ST = 1$ is not a sufficient cause of $BS = 1$ in the context $u^*$ where Suzy is accurate and her rock hits before
Billy’s, but $SH = 1$ is. Recall that $ST = 1$ remains a cause of $BS = 1$ in $(M_{RT}^*, u^*)$. This shows that (part of) a cause is not necessarily part of a sufficient cause.

In the double-prevention example (Example 2.3.4), assuming that Suzy is accurate, $SBT = 1$ (Suzy bombing the target) is a sufficient cause of $TD = 1$ (the target being destroyed). In contrast, Billy pulling the trigger and shooting down Enemy is not, if there are contexts where Suzy does not go up, and thus does not destroy the target. More generally, a but-for cause at the beginning of a long causal chain is unlikely to be a sufficient cause.

The definition of actual causality (Definition 2.2.1) focuses on the actual context; the set of possible contexts plays no role. By way of contrast, the set of contexts plays a major role in the definition of sufficient causality. The use of SC3 makes sufficient causality quite sensitive to the choice of the set of possible contexts. It also may make it an unreasonably strong requirement in some cases. Do we want to require SC3 to hold even in some extremely unlikely contexts? A sensible way to weaken SC3 would be to add probability to the picture, especially if we have a probability on contexts, as in Section 2.5. Rather than requiring SC3 to hold for all contexts, we could then consider the probability of the set of contexts for which it holds. That is, we can take $\vec{X} = \vec{x}$ to be a sufficient cause of $\varphi$ with probability $\alpha$ in $(M, \vec{u})$ if SC1, SC2, and SC4 hold and the set of contexts for which SC3 holds has probability at least $\alpha$.

Notice that with this change, there is a tradeoff between minimality and probability of sufficiency. Consider Suzy throwing rocks. Suppose that, although Suzy is quite accurate, there is a (quite unlikely) context where there are high winds, so Suzy misses the bottle. In this model, a sufficient cause for the bottle shattering is “Suzy throws and there are no high winds”. Suzy throwing is not by itself sufficient. But it is sufficient in all contexts but the one in which there are high winds. If the context with high winds has probability $.1$, then Suzy’s throw is a sufficient cause of the bottle shattering with probability $.9$.

This tradeoff between minimality and probability of sufficiency comes up again in the context of the definition of explanation, considered in Section 7.1, so I do not dwell on it here. But thinking in terms of probability provides some insight into whether not performing an action should be treated as a sufficient cause. Consider Example 2.3.5, where Billy is hospitalized. Besides Billy’s doctor, there are other doctors at the hospital who could, in principle, treat Billy. But they are unlikely to do so. Indeed, they will not even check up on Billy unless they are told to; they presumably have other priorities.

Now consider a context where in fact no doctor treats Billy. In that case, Billy’s doctor not treating Billy on Monday is a sufficient cause for Billy feeling sick on Tuesday with high probability because, with high probability, no other doctor would treat him either. However, another doctor not treating Billy has only a low probability of being a sufficient cause for Billy feeling sick on Tuesday because, with high probability, Billy’s doctor does treat him. This intuition is similar in spirit to the way normality considerations are brought to bear on this example in Chapter 3 (see Example 3.2.2).

For the most part, I do not focus on sufficient causality in this book. I think that notions such as normality, blame, and explanation capture many of the same intuitions as sufficient causality in an arguably better way. However, it is worth keeping sufficient causality in mind when we consider these other notions.
2.7 Causality in Nonrecursive Models

Up to now, I have considered causality only in recursive (i.e., acyclic) causal models. Recursive models are also my focus throughout the rest of the book. However, there are examples that arguably involve nonrecursive models. For example, the connection between pressure and volume is given by Boyle’s Law, which says that $PV = k$: the product of pressure and volume is a constant (if the temperature remains constant). Thus, increasing the pressure decreases the volume, and decreasing the volume increases the pressure. There is no direction of causality in this equation. However, in any particular setting, we are typically manipulating either the pressure or the volume, so there is (in that setting) a “direction” of causality.

Perhaps a more interesting example is one of collusion. Imagine two arsonists who agree to burn down the forest, when only one match suffices to burn down the forest. (So we are essentially in the disjunctive model.) However, the arsonists make it clear to each other that each will not drop a lighted match unless the other does. At the critical moment, they look in each others’ eyes, and both drop their lit matches, resulting in the forest burning down. In this setting, it is plausible that each arsonist caused the other to throw the match, and together they caused the forest to burn down. That is, if $MD_i$ stands for “arsonist $i$ drops a lit match”, for $i = 1, 2$, and $U$ is an exogenous variable that determines whether the arsonists initially intend to drop the match, we can take the equation for $MD_1$ to be such that $MD_1 = 1$ if and only if $U = 1$ and $MD_2 = 1$, and similarly for $MD_2$. This model is not recursive; in all contexts, $MD_1$ depends on $MD_2$ and $MD_2$ depends on $MD_1$. In this section, I show how the HP definition(s) can be extended to deal with such 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, this means that a context no longer necessarily determines the values of the endogenous variables. Earlier, I identified a primitive event such as $X = x$ with the basic causal formula $[](X = x)$, that is, a formula of the form $[y_1 \leftarrow y_1, \ldots, y_k \leftarrow y_k](X = x)$ with $k = 0$. In recursive causal models, where there is a unique solution to the equations in a given context, we can define $[](X = x)$ to be true in $(M, \vec{u})$ if $X = x$ is the unique solution to the equations in context $\vec{u}$. It seems reasonable to identify $[](X = x)$ with $X = x$ in this case. But it is not so reasonable if there may be several solutions to the equations or none.

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 $X = x$ to be relative not just to a context but to a complete assignment: a complete description $(\vec{u}, \vec{v})$ of the values of both the exogenous and the endogenous variables. As before, the context $\vec{u}$ assigns a value to all the exogenous variables; $\vec{v}$ assigns a value to all the endogenous variables. Now define $(M, \vec{u}, \vec{v}) \models X = x$ if $X$ has value $x$ in $\vec{v}$. Since the truth of $X = x$ depends on just $\vec{v}$, not $\vec{u}$, I sometimes write $(M, \vec{v}) \models X = x$.

This definition can be extended to Boolean combinations of primitive events in the standard way. Define $(M, \vec{u}, \vec{v}) \models [\vec{Y} \leftarrow \vec{g}](X = x)$ iff $(M, \vec{v}') \models \varphi$ for all solutions $(\vec{u}, \vec{v}')$ to the equations in $M_{\vec{Y} \leftarrow \vec{g}}$. Since the truth of $[\vec{Y} \leftarrow \vec{g}](X = x)$ depends only on the setting $\vec{u}$ of the exogenous variables, and not on $\vec{v}$, I write $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{g}](X = x)$ to emphasize this.

With these definitions in hand, it is easy to extend the HP definition of causality to arbitrary models. Causality is now defined with respect to a tuple $(M, \vec{u}, \vec{v})$ because we need to know...
2.7 Causality in Nonrecursive Models

the values of the endogenous variables to determine the truth of some formulas. That is, we now talk about $\vec{X} = \vec{x}$ being an actual cause of $\varphi$ in $(M, \vec{u}, \vec{v})$. As before, we have conditions AC1–AC3. AC1 and AC3 remain unchanged. AC2 depends on whether we want to consider the original, updated, or modified definition. For the updated definition, we have:

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

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

(b) If $\vec{z}'$ is such that $(M, \vec{u}, \vec{v}) \models \vec{Z} = \vec{z}'$, then, for all subsets $\vec{W}'$ of $\vec{W}$ and $\vec{Z}'$ of $\vec{Z} - \vec{X}$, $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}'] \varphi$.

In recursive models, the formula $\neg [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}'] \varphi$ in the version of AC2(a) above is equivalent to $[\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}'] \neg \varphi$, the formula used in the original formulation of AC2(a). Given a recursive model $M$, there is a unique solution to the equations in $M_{\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}'}$; $\neg \varphi$ holds in that solution iff $\varphi$ does not hold. However, with nonrecursive models, there may be several solutions; AC2(a) holds if $\varphi$ does not hold in at least one of them. By way of contrast, for AC2(b) to hold, $\varphi$ must stay true in all solutions to the equations in $M_{\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}'}$ in context $\vec{u}$. AC2(a) already says that there is some contingency in which setting $\vec{X}$ to $\vec{x}$ is necessary to bring about $\varphi$ (since setting $\vec{X}$ to something other than $\vec{x}$ in that contingency may result in $\varphi$ no longer holding). Requiring $\neg \varphi$ to hold in only one solution to the equations seems in the spirit of the necessity requirement. However, AC2(b) says that setting $\vec{X}$ to $\vec{x}$ is sufficient to bring about $\varphi$ in all relevant settings, so it makes sense that we want $\varphi$ to hold in all solutions. That’s part of the intuition for sufficiency. Clearly, in the recursive case, where there is only one solution, this definition agrees with the definition given earlier.

Analogous changes are required to extend the original and modified HP definition to the nonrecursive case. For AC2(b), we require only that for all subsets $\vec{Z}$ of $\vec{Z} - \vec{X}$, we have $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}, \vec{Z} \leftarrow \vec{z}'] \varphi$, and do not require this to hold for all subsets $\vec{W}'$ of $\vec{W}$; for AC2(a), we require that $\vec{w}$ consist of the values of the variables in $\vec{W}$ in the actual context. Again, these definitions agree with the earlier definitions in the recursive case.

Note that, with these definitions, all definitions agree that $MD_1 = 1$ and $MD_2 = 1$ are causes of the fire in the collusive disjunctive scenario discussed above. This is perhaps most interesting in the case of the modified HP definition, where in the original disjunctive scenario, the cause was the conjunction $MD = 1 \land L = 1$. However, in this case, each arsonist is a but-for cause of the fire. Although it takes only one match to start the fire, if arsonist 1 does not drop his match, then arsonist 2 won’t drop his match either, so there will not be a fire. The key point is that the arsonists do not act independently here, whereas in the original story, the arsonist and lightning were viewed as independent.

Interestingly, we can make even the collusion story recursive. Suppose that we add a variable $EC$ for eye contact (allowing for the possibility that the arsonists avoid each other’s eyes), where the equations are such that each drops the match only if they do indeed have eye contact. Now we are back at a model where each arsonist is (separately) a cause of the fire (as is the eye contact) according to the original and updated HP definition, and $MD_1 = 1 \land MD_2 = 1$ is a cause according to the modified HP definition. We can debate which is the
“better” model here. But I have no reason to believe that we can always replace a nonrecursive model by a recursive model and still describe essentially the same situation.

Finally, it is worth noting that causality is not necessarily asymmetric in nonrecursive models; we can have $X = x$ being a cause of $Y = y$ and $Y = y$ being a cause of $X = x$. For example, in the model of the collusive story above (without the variable $EC$), $MD_1 = 1$ is a cause of $MD_2 = 1$ and $MD_2 = 1$ is a cause of $MD_1 = 1$. It should not be so surprising that we can have such symmetric causality relations in nonrecursive models. However, it is easy to see that causality is asymmetric in recursive models, according to all the variants of the HP definition. AC2(a) and AC2($a^m$) guarantee that if $X = x$ is a cause of $Y = y$ in $(M, u)$, then $Y$ depends on $X$ in context $u$. Similarly, $X$ must depend on $Y$ if $Y = y$ is a cause of $X = x$ in $(M, u)$. This cannot be the case in a recursive model.

Although the definition of causality in nonrecursive models given here seems like the most natural generalization of the definition for recursive models, I do not have convincing examples suggesting that this is the “right” definition, nor do I have examples showing that this definition is problematic in some ways. This is because all the standard examples are most naturally modeled using recursive models. It would be useful to have more examples of scenarios that are best captured by nonrecursive models for which we have reasonable intuitions regarding the ascription of causality.

2.8 AC2($b^o$) vs. AC2($b^u$)

In this section, I consider more carefully AC2($b^o$) and AC2($b^u$), to give the reader a sense of the subtle differences between original and updated HP definition. I start with the example that originally motivated replacing AC2($b^o$) by AC2($b^u$).

**Example 2.8.1** Suppose that a prisoner dies either if $A$ loads B’s gun and $B$ shoots or if $C$ loads and shoots his gun. Taking $D$ to represent the prisoner’s death and making the obvious assumptions about the meaning of the variables, we have that $D = (A \land B) \lor C$. (Note that here I am identifying the binary variables $A$, $B$, $C$, and $D$ with primitive propositions in propositional logic, as I said earlier I would. I could have also written this as $D = \max(\min(A, B), C)$.) Suppose that in the actual context $u$, $A$ loads B’s gun, $B$ does not shoot, but $C$ does load and shoot his gun, so the prisoner dies. That is, $A = 1$, $B = 0$, and $C = 1$. Clearly $C = 1$ is a cause of $D = 1$. We would not want to say that $A = 1$ is a cause of $D = 1$, given that $B$ did not shoot (i.e., given that $B = 0$). However, suppose that we take the obvious model with the variables $A$, $B$, $C$, $D$. With AC2($b^o$), $A = 1$ 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($b^o$) hold for this contingency, since $(M, u) \models [A \leftarrow 0, B \leftarrow 1, C \leftarrow 0](D = 0)$, whereas $(M, u) \models [A \leftarrow 1, B \leftarrow 1, C \leftarrow 0](D = 1)$. Thus, according to the original HP definition, $A = 1$ is a cause of $D = 1$. However, AC2($b^u$) fails in this case because $(M, u) \models [A \leftarrow 1, C \leftarrow 0](D = 0)$. The key point is that AC2($b^u$) says that for $A = 1$ to be a cause of $D = 1$, it must be the case that $D = 1$ even if only some of the values in $\vec{w}$ are set to their values $\vec{w}$. In this case, by setting only $A$ to 1 and leaving $B$ unset, $B$ takes on its original value of 0, in which case $D = 0$. AC2($b^o$) does not consider this case.
Note that the modified HP definition also does not call $A = 1$ a cause of $D = 1$. This follows from Theorem 2.2.3 and the observations above; it can also be seen easily by observing that if the values of any subset of variables are fixed at their actual values in context $u$, setting $A = 0$ will not make $D = 0$.

Although it seems clear that we do not want $A = 1$ to be a cause of $D = 1$, the situation for the original HP definition is not as bleak as it may appear. We can deal with this example using the original HP definition if we add extra variables, as we did with the Billy-Suzy rock-throwing example. Specifically, suppose that we add a variable $B'$ for “$B$ shoots a loaded gun”, where $B' = A \land B$ and $D = B' \lor C$. This slight change prevents $A = 1$ from being a cause of $D = 1$, even according to the original HP definition. I leave it to the reader to check that any attempt to now declare $A = 1$ a cause of $D = 1$ according to the original HP definition would have to put $B'$ into $\vec{Z}$. However, because $B' = 0$ in the actual world, AC2(b°) does not hold. As I show in Section 4.3, this approach to dealing with the problem generalizes. There is a sense in which, by adding enough extra variables, we can always use AC2(b°) instead of AC2(bu) to get equivalent judgments of causality.

There are advantages to using AC2(b°). For one thing, it follows from Theorem 2.2.3(d) that with AC2(b°), causes are always single conjuncts. This is not the case with AC2(bu), as the following example shows.

Example 2.8.2  $A$ votes for a candidate. $A$’s vote is recorded in two optical scanners $B$ and $C$. $D$ collects the output of the scanners; $D'$ records whether just scanner $B$ records a vote for the candidate. The candidate wins (i.e., $WIN = 1$) if any of $A$, $D$, or $D'$ is 1. The value of $A$ is determined by the exogenous variable. The following structural equations characterize the value of the remaining variables:

- $B = A$;
- $C = A$;
- $D = B \land C$;
- $D' = B \land \neg A$; and
- $WIN = A \lor D \lor D'$.

Call this causal model $M_V$. Roughly speaking, $D'$ acts like BH in the rock-throwing example as modeled by $M_{RT}'$. The causal network for $M_V$ is shown in Figure 2.9.

In the actual context $u$, $A = 1$, so $B = C = D = WIN = 1$ and $D' = 0$. I claim that $B = 1 \land C = 1$ is a cause of $WIN = 1$ in $(M_V, u)$ according to the updated HP definition (which means that, by AC3, neither $B = 1$ nor $C = 1$ is cause of $WIN = 1$ in $(M_V, u)$). To see this, first observe that AC1 clearly holds. For AC2, consider the witness where $\hat{W} = \{A\}$ (so $\hat{Z} = \{B, C, D, D', WIN\}$) and $w = 0$ (so we are considering the contingency where $A = 0$). Clearly, $(M_V, u) \models [A \leftarrow 0, B \leftarrow 0, C \leftarrow 0](WIN = 0)$, so AC2(a) holds, and $(M_V, u) \models [A \leftarrow 0, B \leftarrow 1, C \leftarrow 1](WIN = 1)$. Moreover, $(M_V, u) \models [B \leftarrow 1, C \leftarrow 1](WIN = 1)$, and $WIN = 1$ continues to hold even if $D$ is set to 1 and/or $D'$ is set to 0 (their values in $(M_V, u)$). Thus, AC2 holds.
It remains to show that AC3 holds and, in particular, that neither \( B = 1 \) nor \( C = 1 \) is a cause of \( WIN = 1 \) in \((M_V, u)\) according to the original HP definition. The argument is essentially the same for both \( B = 1 \) and \( C = 1 \), so I just show it for \( B = 1 \). Roughly speaking, \( B = 1 \) does not satisfy AC2(a) and AC2(b) for the same reason that \( BT = 1 \) does not satisfy it in the rock-throwing example. For suppose that \( B = 1 \) satisfies AC2(a) and AC2(b). Then we would have to have \( A \in \bar{W} \), and we would need to consider the contingency where \( A = 0 \) (otherwise \( WIN = 1 \) no matter how we set \( B \)). Now we need to consider two cases: \( D' \in \bar{W} \) and \( D' \in \bar{Z} \). If \( D' \in \bar{W} \), then if we set \( D' = 0 \), we have \((M_V, u) \models [A \leftarrow 0, B \leftarrow 1, D' \leftarrow 0](WIN = 0)\), so AC2(b) fails (no matter whether \( C \) and \( D \) are in \( \bar{W} \) or \( \bar{Z} \)). And if we set \( D' = 1 \), then AC2(a) fails, since \((M_V, u) \models [A \leftarrow 0, B \leftarrow 0, D' \leftarrow 1](WIN = 1)\). Now if \( D' \in \bar{Z} \), note that \((M_V, u) \models D' = 0.\) Since \((M_V, u) \models [A \leftarrow 0, B \leftarrow 1, D' \leftarrow 0](WIN = 0)\), again AC2(b) fails (no matter whether \( C \) or \( D \) are in \( \bar{W} \) or \( \bar{Z} \)). Thus, \( B = 1 \) is not a cause of \( WIN = 1 \) in \((M_V, u)\) according to the updated HP definition.

By way of contrast, \( B = 1 \) and \( C = 1 \) are causes of \( WIN = 1 \) in \((M_V, u)\) according to the original HP definition (as we would expect from Theorem 2.2.3). Although \( B = 1 \) and \( C = 1 \) do not satisfy AC2(b), they do satisfy AC2(b'). To see that \( B = 1 \) satisfies AC2(b'), take \( \bar{Z} = \{B, D, WIN\} \) and consider the witness where \( A = 0, C = 1, \) and \( D' = 0.\) Clearly we have both that \((M_V, u) \models [B \leftarrow 0, A \leftarrow 0, C \leftarrow 1, D' \leftarrow 0](WIN = 0)\) and that \((M_V, u) \models [B \leftarrow 1, A \leftarrow 0, C \leftarrow 1, D' \leftarrow 0](WIN = 1)\), and this continues to hold if \( D \) is set to 1 (its original value). A similar argument shows that \( C = 1 \) is a cause.

Interestingly, none of \( B = 1, C = 1, \) or \( B = 1 \land C = 1 \) is a cause of \( WIN = 1 \) in \((M_V, u)\) according to the modified HP definition. This follows from Theorem 2.2.3; if \( B = 1 \) were part of a cause according to the modified HP definition, then it would be a cause according to the updated HP definition, and similarly for \( C = 1 \). It can also be seen directly: setting \( B = C = 0 \) has no effect on \( WIN \) if \( A = 1.\) The only cause of \( WIN = 1 \) in \((M_V, u)\) according to the modified HP definition is \( A = 1 \) (which, of course, is also a cause according to the original and updated definitions).

The fact that causes may not be singletons has implications for the difficulty of determining whether \( A \) is an actual cause of \( B \) (see Section 5.3). Although this may suggest that we should
then use the original HP definition, adding extra variables if needed, this conclusion is not necessarily warranted either. For one thing, it may not be obvious when modeling a problem that extra variables are needed. Moreover, the advantages of lower complexity may be lost if sufficiently many additional variables need to be added. The modified HP definition has the benefits of both typically giving the “right” answer without the need for extra variables (but see Section 4.3), while having lower complexity than either the original and modified HP definition (again, see Section 5.3).

I conclude this section with one more example that I hope clarifies some of the details of AC2(b°) (and AC2(b′)). Recall that AC2(b′) requires that \( \varphi \) remain true if \( \vec{X} \) is set to \( \vec{x} \), even if only a subset of the variables in \( \vec{W} \) are set to their values in \( \vec{w} \) and all the variables in an arbitrary subset \( \vec{Z}' \) of \( \vec{Z} \) are set to their original values \( z^* \) (i.e., the values they had in the original context, where \( \vec{X} = \vec{x} \)). Since I have viewed the variables in \( \vec{Z} \) as being on the causal path to \( \varphi \) (although that is not quite right; see Section 2.9), we might hope that not only is setting \( \vec{X} \) to \( \vec{x} \) sufficient to make \( \varphi \) true, but it is also enough to force all the variables in \( \vec{Z} \) to their original values. Indeed, there is a definition of actual causality that makes this requirement (see the bibliographic notes). Formally, we might hope that if \((M, \vec{u}) \models \vec{Z} = \vec{z}^*\), then the following holds:

\[
(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}] (Z = z^*) \quad \text{for all } Z \in \vec{Z} \text{ and all } \vec{W}' \subseteq \vec{W}.
\]  

(2.1)

It is not hard to show (see Lemma 2.10.2) that, in the presence of (2.1), AC2(b°) can be simplified to

\[
(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}] \varphi;
\]

there is no need to include the subsets \( Z' \). Although (2.1) holds in many examples, it seems unreasonable to require it in general, as the following example shows.

**Example 2.8.3** 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 situation where either a match or a lightning strike suffices to bring about a forest fire. 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 part of a cause of \( P = 1 \) according to the original and updated HP definition. However, suppose that the story is modified slightly. Now assume that there is a voting machine that tabulates the votes. Let \( T \) represent the total number of votes recorded by the machine. Clearly, \( T = V_1 + V_2 \) and \( P = 1 \) iff \( T \geq 1 \). The causal network in Figure 2.10 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 \( P = 1 \) according to the original and updated HP definition. Consider \( V_1 = 1 \). Take \( \vec{Z} = \{V_1, T, P\} \) and \( \vec{W} = V_2 \), and consider the contingency \( V_2 = 0 \). With this witness, \( P \) is counterfactually dependent on \( V_1 \), so AC2(a) holds. To check that this contingency satisfies AC2(b°) (and hence also AC2(b°)), note that setting \( V_1 \) to 1 and \( V_2 \) to 0 results in \( P = 1 \), even if \( T \) is set to 2 (its current value). However, (2.1) does not hold here: \( T \) does not retain its original value of 2 when \( V_1 = 1 \) and \( V_2 = 0 \).
Chapter 2. The HP Definition of Causality

Since, in general, one can always imagine that a change in one variable produces some feeble change in another, it seems unreasonable to insist on the variables in $\vec{Z}$ remaining constant; $AC2(b^o)$ and $AC2(b^u)$ require merely that changes in $\vec{Z}$ not affect $\varphi$. □

2.9 Causal Paths

There is an intuition that causality travels along a path: $A$ causes $B$, which causes $C$, which causes $D$, so there is a path from $A$ to $B$ to $C$ to $D$. And, indeed, many accounts of causality explicitly make use of causal paths. In the original and updated HP definition, $AC2$ involves a partition of the endogenous variables into two sets, $\vec{Z}$ and $\vec{W}$. In principle, there are no constraints on what partition to use, other than the requirement that $\vec{X}$ be a subset of $\vec{Z}$. However, when introducing $AC2(a)$, I suggested that $\vec{Z}$ can be thought of as consisting of the variables on the causal path from $\vec{X}$ to $\varphi$. (Recall that the notion of a causal path is defined formally in Section 2.4.) As I said, this intuition is not quite true for the updated HP definition; however, it is true for the original HP definition. In this section, I examine the role of causal paths in $AC2$. In showing that $\vec{X}$ is a cause of $\varphi$, I would like to show that we can take $\vec{Z}$ (or $\vec{V} - \vec{W}$ in the case of the modified HP definition) to consist only of variables that lie on a causal path from some variable in $\vec{X}$ to some variable in $\varphi$. The following example shows that this is not the case in general with the updated HP definition.

Example 2.9.1 Consider the following variant of the disjunctive forest-fire scenario. Now there is a second arsonist; he will drop a lit match exactly if the first one does. Thus, we add a new variable $MD'$ with the equation $MD' = MD$. Moreover, $FF$ is no longer a binary variable; it has three possible values: 0, 1, and 2. We have that $FF = 0$ if $L = MD = MD' = 0$; $FF = 1$ if either $L = 1$ and $MD = MD'$ or $MD = MD' = 1$; and $FF = 2$ if $MD \neq MD'$. We can suppose that the forest burns asymmetrically ($FF = 2$) if exactly one of the arsonists drops a lit match (which does not happen under normal circumstances); one side of the forest is more damaged than the other (even if the lightning strikes). Call the resulting causal model $M'_{FF}$. Figure 2.11 shows the causal network corresponding to $M'_{FF}$. 

Figure 2.10: A more refined voting scenario.
In the context where $L = MD = 1$, $L = 1$ is a cause of $FF = 1$ according to the updated HP definition. Note that there is only one causal path from $L$ to $FF$: the path consisting of just the variables $L$ and $FF$. But to show that $L = 1$ is a cause of $FF = 1$ according to the updated HP definition, we need to take the witness to be $\{MD\}, 0, 0$; in particular, $MD'$ is part of $\vec{Z}$, not $\vec{W}$.

To see that this witness works, note that $(M_{FF}', u) \models [L \leftarrow 0, MD \leftarrow 0](FF = 0)$, so AC2(a) holds. Since we have both $(M_{FF}', u) \models [L \leftarrow 1, MD \leftarrow 0](FF = 1)$ and $(M_{FF}', u) \models [L \leftarrow 1](FF = 1)$, AC2(b) also holds. However, suppose that we wanted to find a witness where the set $\vec{W} = \{MD, MD'\}$ (so that $\vec{Z}$ can be the unique causal path from $L$ to $FF$). What do we set $MD$ and $MD'$ to in the witness? At least one of them must be 0, so that when $L$ is set to 0, $FF \neq 1$. Setting $MD' = 0$ does not work: $(M_{FF}', u) \models [L \leftarrow 1, MD' \leftarrow 0](FF = 2)$, so AC2(b) does not hold. And setting $MD = 0$ and $MD' = 1$ does not work either, for then $FF = 2$, independent of the value of $L$, so again AC2(b) does not hold.

By way of contrast, we can take $(\{MD, MD'\}, \vec{0}, 0)$ to be a witness to $L = 1$ being a cause according to the original HP definition; AC2(b) holds with this witness because $(M_{FF}', u) \models [L \leftarrow 1, MD \leftarrow 0, MD' \leftarrow 0](FF = 1)$.

Finally, perhaps disconcertingly, $L = 1$ is not even part of a cause of $FF = 1$ in $(M, u)$ according to the modified HP definition. That is because $MD = 1$ is now a cause; setting $MD$ to 0 while keeping $MD'$ fixed at 1 results in $FF$ being 2. Thus, unlike the original forest-fire example, $L = 1 \land MD = 1$ is no longer a cause of $FF = 1$ according to the modified HP definition. This problem disappears once we take normality into account; see Example 3.2.5.

For the original HP definition, we can take the set $\vec{Z}$ in AC2, not just in this example, but in general, to consist only of variables that lie on a causal path from a variable in $\vec{X}$ to a variable in $\varphi$. Interestingly, it is also true of the modified HP definition, except in that case we have to talk about $\varphi - \vec{W}$ rather than $\vec{Z}$ because $\varphi - \vec{W}$ is the analogue of $\vec{Z}$ in the modified HP definition.

**Proposition 2.9.2** If $\vec{X} = \vec{x}$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the original or modified HP definition, then there exists a witness $(\vec{W}, \vec{w}, \vec{x}')$ to this such that every variable $Z \in \varphi - \vec{W}$ lies on a causal path in $(M, \vec{u})$ from some variable in $\vec{X}$ to some variable in $\varphi$.

For definiteness, I take $X$ to be a variable “in $\varphi$” if, syntactically, $X$ appears in $\varphi$, even if $\varphi$ is equivalent to a formula that does not involve $X$. For example, if $Y$ is a binary variable,
I take $Y$ to appear in $X = 1 \land (Y = 1 \lor Y = 1)$, although this formula is equivalent to $X = 1$. (The proposition is actually true even if I assume that $Y$ does not appear in $X = 1 \land (Y = 1 \lor Y = 1)$.) The proof of the proposition can be found in Section 2.10.3.

### 2.10 Proofs

In this section, I prove some of the results stated earlier. This section can be skipped without loss of continuity.

#### 2.10.1 Proof of Theorem 2.2.3

I repeat the statements of the theorem here (and in the later sections) for the reader's convenience.

**Theorem 2.2.3:**

(a) If $X = x$ is part of a cause of $\varphi$ in $(M, \vec{u})$ according to the modified HP definition, then $X = x$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the original HP definition.

(b) If $X = x$ is part of a cause of $\varphi$ in $(M, \vec{u})$ according to the modified HP definition, then $X = x$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the updated HP definition.

(c) If $X = x$ is part of a cause of $\varphi$ in $(M, \vec{u})$ according to the updated HP definition, then $X = x$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the original HP definition.

(d) If $\vec{X} = \vec{x}$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the original HP definition, then $|\vec{X}| = 1$ (i.e., $\vec{X}$ is a singleton).

**Proof:** For part (a), suppose that $X = x$ is part of a cause of $\varphi$ in $(M, \vec{u})$ according to the modified HP definition, so that there is a cause $\vec{X} = \vec{x}$ such that $X = x$ is one of its conjuncts. I claim that $X = x$ is a cause of $\varphi$ according to the original HP definition. By definition, there must exist a value $\vec{x}' \in \mathcal{R}(\vec{X})$ and a set $\vec{W} \subseteq \mathcal{V} - \vec{X}$ such that if $(M, \vec{u}) \models \vec{W} = \vec{w}^*$, then $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}^*] \neg \varphi$. Moreover, $\vec{X}$ is minimal.

To show that $X = x$ is a cause according to the original HP definition, we must find an appropriate witness. If $\vec{X} = \{X\}$, then it is immediate that $(\vec{W}, \vec{w}^*, x')$ is a witness. If $|\vec{X}| > 1$, suppose without loss of generality that $\vec{X} = (X_1, \ldots, X_n)$, and $X = X_1$. In general, if $\vec{Y}$ is a vector, I write $\vec{Y}_{-1}$ to denote all components of the vector except the first one, so that $\vec{X}_{-1} = (X_2, \ldots, X_n)$. I want to show that $X_1 = x_1$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the original HP definition. Clearly, $(M, \vec{u}) \models (X_1 = x_1) \land \varphi$, since $\vec{X} = \vec{x}$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the modified HP definition, so AC1 holds. The obvious candidate for a witness for AC2(a) is $(\vec{X}_{-1} \cdot \vec{W}, \vec{x}'_{-1} \cdot \vec{w}^*, x'_1)$, where $\cdot$ is the operator that concatenates two vectors so that, for example, $(1, 3) \cdot (2) = (1, 3, 2)$; that is, roughly speaking, we move $X_2, \ldots, X_n$ into $\vec{W}$. This satisfies AC2(a), since $(M, \vec{u}) \models [X_1 \leftarrow x'_1, \vec{X}_{-1} \leftarrow \vec{x}'_{-1}, \vec{W} \leftarrow \vec{w}^*] \neg \varphi$ by assumption. AC3 trivially holds for $X_1 = x_1$, so it remains to deal with AC2(b'). Suppose, by way of contradiction, that $(M, \vec{u}) \models [X_1 \leftarrow x_1, \vec{X}_{-1} \leftarrow \vec{x}'_{-1}, \vec{W} \leftarrow \vec{w}^*] \neg \varphi$. This means...
that $\bar{X}_{-1} \leftarrow \bar{x}_{-1}$ satisfies AC2($a(m)$), showing that AC3 (more precisely, the version of AC3 appropriate for the modified HP definition) is violated (taking $((X_1) \cdot \bar{W}, (x_1) \cdot \bar{w}^*, \bar{x}_{-1}')$ as the witness), and $\bar{X} \leftarrow \bar{x}$ is not a cause of $\varphi$ in $(M, \bar{u})$ according to the modified HP definition, a contradiction. Thus, $(M, \bar{u}) \models [X_1 \leftarrow x_1, \bar{X}_{-1} \leftarrow \bar{x}_{-1}', \bar{W} \leftarrow \bar{w}^*] \varphi$.

This does not yet show that AC2($b^o$) holds; there might be some subset $\bar{Z}'$ of variables in $\mathcal{V} - \bar{X}_{-1} \cup \bar{W}$ that change value when $\bar{W}$ is set to $\bar{w}^*$ and $\bar{X}_{-1}$ is set to $\bar{x}_{-1}$, and when these variables are set to their original values in $(M, \bar{u})$, $\varphi$ does not hold, thus violating AC2($b^u$). In more detail, suppose that there exist a set $\bar{Z}' = (Z_1, \ldots, Z_k) \subseteq \bar{Z}$ and values $z^*_j$ for each variable $Z_j \in \bar{Z}'$ such that (i) $(M, \bar{u}) \models Z_j = z^*_j$ and (ii) $(M, \bar{u}) \models [X_1 \leftarrow x_1, \bar{X}_{-1} \leftarrow \bar{x}_{-1}, \bar{W} \leftarrow \bar{w}^*, \bar{Z}' = \bar{z}^*] \neg \varphi$. But then $\bar{X} = \bar{x}$ is not a cause of $\varphi$ in $(M, \bar{u})$ according to the modified HP definition. Condition (ii) shows that AC2($a(m)$) is satisfied for $\bar{X}_{-1}$, taking $((X_1) \cdot \bar{W} \cdot \bar{Z}', (x_1) \cdot \bar{w}^* \cdot \bar{z}^*, \bar{x}_{-1}')$ as the witness, so again AC3 is violated. It follows that AC2($b^o$) holds, completing the proof.

The proof of part (b) is similar in spirit. Indeed, we just need to show one more thing. For AC2($b^u$), we must show that if $\bar{X}' \subseteq \bar{X}_{-1}$, $\bar{W}' \subseteq \bar{W}$, and $\bar{Z}' \subseteq \bar{Z}$, then

$$(M, \bar{u}) \models [X_1 \leftarrow x_1, \bar{X}' \leftarrow \bar{x}'_{-1}, \bar{W}' \leftarrow \bar{w}^*, \bar{Z}' \leftarrow \bar{z}^*] \varphi. \tag{2.2}$$

(Here I am using the abuse of notation that I referred to in Section 2.2.2, where if $\bar{X}' \subseteq \bar{X}$ and $\bar{x} \in \mathcal{R}(\bar{X})$, I write $\bar{X}' \leftarrow \bar{x}$, with the intention that the components of $\bar{x}'$ not included in $\bar{X}'$ are ignored.) It follows easily from AC1 that (2.2) holds if $\bar{X}' = \emptyset$. And if (2.2) does not hold for some strict nonempty subset $\bar{X}'$ of $\bar{X}_{-1}$, then $\bar{X} = \bar{x}$ is not a cause of $\varphi$ according to the modified HP definition because AC3 does not hold; AC2($a(m)$) is satisfied for $\bar{X}'$.

For part (c), the proof is again similar in spirit. Indeed, the proof is identical up to the point where we must show that AC2($b^o$) holds. Now if $\bar{Z}' \subseteq \bar{Z}$ and $(M, \bar{u}) \models [X_1 \leftarrow x_1, \bar{X}_{-1} \leftarrow \bar{x}''_{-1}, \bar{W} \leftarrow \bar{w}, \bar{Z}' \leftarrow \bar{z}^*] \neg \varphi$, then $\bar{X}_{-1} \leftarrow \bar{x}_{-1}$ satisfies AC2($a$) (taking $((X_1) \cdot \bar{W} \cdot \bar{Z}'', (x_1) \cdot \bar{w} \cdot \bar{z}^*, \bar{x}_{-1}')$ as the witness). It also satisfies AC2($b^u$), since $\bar{X} = \bar{x}$ satisfies AC2($b^u$), by assumption. Thus, the version of AC3 appropriate for the updated HP definition is violated.

Finally, for part (d), the proof is yet again similar in spirit. Suppose that $\bar{X} = \bar{x}$ is a cause of $\varphi$ in $(M, \bar{u})$ according to the original HP definition and $|\bar{X}| > 1$. Let $X = x$ be a conjunct of $\bar{X} = \bar{x}$. Again, we can show that $X = x$ is a cause of $\varphi$ in $(M, \bar{u})$ according to the original HP definition, which is exactly what is needed to prove part (d), using essentially the same argument as above. I leave the details to the reader.}

### 2.10.2 Proof of Proposition 2.4.6

In this section, I prove Proposition 2.4.6. I start with a simple lemma that states a key (and obvious!) property of causal paths: if there is no causal path from $X$ to $Y$, then changing the value of $X$ cannot change the value of $Y$. This fact is made precise in the following lemma. Although it is intuitively obvious, proving it carefully requires a little bit of work. I translate the proof into statisticians’ notation as well.
Lemma 2.10.1 If \( Y \) and all the variables in \( \vec{X} \) are endogenous, \( Y \notin \vec{X} \), and there is no causal path in \((M, \vec{u})\) from a variable in \( \vec{X} \) to \( Y \), then for all sets \( \vec{W} \) of variables disjoint from \( \vec{X} \) and \( Y \), and all settings \( \vec{x} \) and \( \vec{x}' \) for \( \vec{X} \), \( y \) for \( Y \), and \( \vec{w} \) for \( \vec{W} \), we have

\[
(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}](Y = y) \iff (M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}](Y = y)
\]

and

\[
(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}](Y = y) \iff (M, \vec{u}) \models [\vec{W} \leftarrow \vec{w}](Y = y)
\]

(i.e., \( Y_{\vec{x}\vec{w}}(\vec{u}) = y \) iff \( Y_{\vec{x}'\vec{w}}(\vec{u}) = y \) and \( Y_{\vec{x}\vec{w}}(\vec{u}) = y \) iff \( Y_{\vec{w}}(\vec{u}) = y \).

Proof: Define the maximum distance of a variable \( Y \) in a causal setting \((M, \vec{u})\), denoted \( \text{maxdist}(Y) \), to be the length of the longest causal path in \((M, \vec{u})\) from an exogenous variable to \( Y \). We prove both parts of the result simultaneously by induction on \( \text{maxdist}(Y) \). If \( \text{maxdist}(Y) = 1 \), then the value of \( Y \) depends only on the values of the exogenous variables, so the result trivially holds. If \( \text{maxdist}(Y) > 1 \), let \( Z_1, \ldots, Z_k \) be the endogenous variables on which \( Y \) depends. These are the parents of \( Y \) in the causal network (i.e., these are exactly the endogenous variables \( Z \) such that there is an edge from \( Z \) to \( Y \) in the causal network). For each \( Z \in \{Z_1, \ldots, Z_k\} \), \( \text{maxdist}(Z) < \text{maxdist}(Y) \): for each causal path in \((M, \vec{u})\) from an exogenous variable to \( Z \), there is a longer path to \( Y \), namely, the one formed by adding the edge from \( Z \) to \( Y \). Moreover, there is no causal path in \((M, \vec{u})\) from a variable in \( \vec{X} \) to any of \( Z_1, \ldots, Z_k \) nor is any of \( Z_1, \ldots, Z_k \) in \( \vec{X} \) (for otherwise there would be a causal path in \((M, \vec{u})\) from a variable in \( \vec{X} \) to \( Y \), contradicting the assumption of the lemma). Thus, the inductive hypothesis holds for each of \( Z_1, \ldots, Z_k \). Since the value of each of \( Z_1, \ldots, Z_k \) does not change when we change the setting of \( \vec{X} \) from \( \vec{x} \) to \( \vec{x}' \), and the value of \( Y \) depends only on the values of \( Z_1, \ldots, Z_k \) and \( \vec{u} \) (i.e., the values of the exogenous variables), the value of \( Y \) cannot change either. \( \square \)

I can now prove Proposition 2.4.6. I restate it for the reader’s convenience.

Proposition 2.4.6: Suppose that \( X_1 = x_1 \) is a but-for cause of \( X_2 = x_2 \) in the causal setting \((M, \vec{u})\), \( X_2 = x_2 \) is a but-for cause of \( X_3 = x_3 \) in \((M, \vec{u})\), and the following two conditions hold:

(a) for every value \( x_2' \in \mathcal{R}(X_2) \), there exists a value \( x_1' \in \mathcal{R}(X_1) \) such that \((M, \vec{u}) \models [X_1 \leftarrow x_1'](X_2 = x_2') \) (i.e., \((X_2)_{x_1'}(\vec{u}) = x_2')\);

(b) \( X_2 \) is on every causal path in \((M, \vec{u})\) from \( X_1 \) to \( X_3 \).

Then \( X_1 = x_1 \) is a but-for cause of \( X_3 = x_3 \).

Proof: Since \( X_2 = x_2 \) is a but-for cause of \( X_3 = x_3 \) in \((M, \vec{u})\), there exists a value \( x_2' \neq x_2 \) such that \((M, \vec{u}) \models [X_2 \leftarrow x_2'](X_3 \neq x_3') \) (i.e., \((X_3)_{x_2'}(\vec{u}) = x_3')\). Choose \( x_3' \) such that \((M, \vec{u}) \models [X_2 \leftarrow x_2'](X_3 = x_3') \) (i.e., \((X_3)_{x_2'}(\vec{u}) = x_3')\). By assumption, there exists a value \( x_1' \) such that \((M, \vec{u}) \models [X_1 \leftarrow x_1'](X_2 = x_2) \) (i.e., \((X_2)_{x_1'}(\vec{u}) = x_2')\). I claim that \((M, \vec{u}) \models [X_1 \leftarrow x_1'](X_3 = x_3') \) (i.e., \((X_3)_{x_1'}(\vec{u}) = x_3')\). This follows from a more general claim. I show that if \( Y \) is on a causal path from \( X_2 \) to \( X_3 \), then

\[
(M, \vec{u}) \models [X_1 \leftarrow x_1'](Y = y) \iff (M, \vec{u}) \models [X_2 \leftarrow x_2'](Y = y)
\] (2.3)
2.10 Proofs

Define a partial order $\preceq$ on endogenous variables that lie on a causal path from $X_2$ to $X_3$ by taking $Y_1 \prec Y_2$ if $Y_1$ precedes $Y_2$ on some causal path from $X_2$ to $X_3$. Since $M$ is a recursive model, if $Y_1$ and $Y_2$ are distinct variables and $Y_1 \prec Y_2$, we cannot have $Y_2 \prec Y_1$ (otherwise there would be a cycle). I prove (2.3) by induction on the $\prec$ ordering. The least element in this ordering is clearly $X_2$; $X_2$ must come before every other variable on a causal path from $X_2$ to $X_3$. $(M, \vec{u}) \models [X_1 \prec x'_1][X_2 = x'_2]$ (i.e., $(X_2)_{x'_1}(\vec{u}) = x'_2$) by assumption, and clearly $(M, \vec{u}) \models [X_2 \prec x'_2][X_2 = x'_2]$ (i.e., $(X_2)_{x'_2}(\vec{u}) = x'_2$). Thus, (2.3) holds for $X_2$. Thus completes the base case of the induction.

For the inductive step, let $Y$ be a variable that lies on a causal path in $(M, \vec{u})$ from $X_2$ and $X_3$, and suppose that (2.3) holds for all variables $Y'$ such that $Y' \prec Y$. Let $Z_1, \ldots, Z_k$ be the endogenous variables that $Y$ depends on in $M$. For each of these variables $Z_i$, either there is a causal path in $(M, \vec{u})$ from $X_1$ to $Z_i$ or there is not. If there is, then the path from $X_1$ to $Z_i$ can be extended to a directed path $P$ from $X_1$ to $X_3$ by going from $X_1$ to $Z_i$, from $Z_i$ to $Y$, and from $Y$ to $X_3$ (since $Y$ lies on a causal path in $(M, \vec{u})$ from $X_2$ to $X_3$). Since, by assumption, $X_2$ lies on every causal path in $(M, \vec{u})$ from $X_1$ to $X_3$, $X_2$ must lie on $P$. Moreover, $X_2$ must precede $Y$ on $P$. (Proof: Since $Y$ lies on a path $P'$ from $X_2$ to $X_3$, $X_2$ must precede $Y$ on $P'$. If $Y$ precedes $X_2$ on $P$, then there is a cycle, which is a contradiction.) Since $Z_i$ precedes $Y$ on $P$, it follows that $Z_i \prec Y$, so by the inductive hypothesis, $(M, \vec{u}) \models [X_1 \prec x'_1](Z_i = z_i)$ iff $(M, \vec{u}) \models [X_2 \prec x'_2](Z_i = z_i)$ (i.e., $(Z_i)_{x'_1}(\vec{u}) = z_i$ iff $(Z_i)_{x'_2}(\vec{u}) = z_i$).

Now if there is no causal path in $(M, \vec{u})$ from $X_1$ to $Z_i$, then there also cannot be a causal path $P$ from $X_2$ to $Z_i$ (otherwise there would be a causal path in $(M, \vec{u})$ from $X_1$ to $Z_i$ formed by appending $P$ to a causal path from $X_1$ to $X_2$, which must exist because, if not, it easily follows from Lemma 2.10.1 that $X_1 = x_1$ would not be a cause of $X_2 = x_2$). Since there is no causal path in $(M, \vec{u})$ from $X_1$ to $Z_i$, we must have that $(M, \vec{u}) \models [X_1 \prec x'_1](Z_i = z_i)$ iff $(M, \vec{u}) \models [X_2 \prec x'_2](Z_i = z_i)$ (i.e., $(Z_i)_{x'_1}(\vec{u}) = z_i$ iff $(Z_i)_{x'_2}(\vec{u}) = z_i$).

Since the value of $Y$ depends only on the values of $Z_1, \ldots, Z_k$ and $\vec{u}$, and I have just shown that $(M, \vec{u}) \models [X_1 \prec x'_1](Z_1 = z_1 \land \ldots \land Z_k = z_k)$ iff $(M, \vec{u}) \models [X_2 \prec x'_2](Z_1 = z_1 \land \ldots \land Z_k = z_k)$ (i.e., $(Z_1)_{x'_1}(\vec{u}) = z_1$ and $(Z_k)_{x'_2}(\vec{u}) = z_k$) it follows that $(M, \vec{u}) \models [X_1 \prec x'_1](Y = y)$ iff $(M, \vec{u}) \models [X_2 \prec x'_2](Y = y)$ (i.e., $Y_{x'_1}(\vec{u}) = y$ iff $Y_{x'_2}(\vec{u}) = y$).

This completes the proof of the induction step. Since $X_3$ is on a causal path in $(M, \vec{u})$ from $X_2$ to $X_3$, it now follows that $(M, \vec{u}) \models [X_1 \prec x'_1](X_3 = x'_3)$ iff $(M, \vec{u}) \models [X_2 \prec x'_2](X_3 = x'_3)$ (i.e., $(X_3)_{x'_1}(\vec{u}) = x'_3$ iff $(X_3)_{x'_2}(\vec{u}) = x'_3$). Since $(M, \vec{u}) \models [X_2 \prec x'_2](X_3 = x'_3)$ (i.e., $(X_3)_{x'_2}(\vec{u}) = x'_3$) by construction, we have that $(M, \vec{u}) \models [X_1 \prec x'_1][X_3 = x'_3]$ (i.e., $(X_3)_{x'_1}(\vec{u}) = x'_3$), as desired. Thus, $X_1 = x_1$ is a but-for cause for $X_3 = x_3$. □

2.10.3 Proof of Proposition 2.9.2

In this section, I prove Proposition 2.9.2.
Proposition 2.9.2: If \( \vec{X} = \vec{x} \) is a cause of \( \varphi \) in \((M, \vec{u})\) according to the original or modified HP definition, then there exists a witness \((\vec{W}, \vec{w}, \vec{x}')\) to this such that every variable \( Z \in \mathcal{V} - \vec{W} \) lies on a causal path in \((M, \vec{u})\) from some variable in \( \vec{X} \) to some variable in \( \varphi \).

To prove the result, I need to prove a general property of causal models, one sufficiently important to be an axiom in the axiomatization of causal reasoning given in Section 5.4.

Lemma 2.10.2 If \( (M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}](\vec{Y} = \vec{y}) \), then \( (M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}]\varphi \) if and only if \( (M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{Y} \leftarrow \vec{y}]\varphi \).

Proof: In the unique solution to the equations when \( \vec{X} \) is set to \( \vec{x} \), \( \vec{Y} = \vec{y} \). So the unique solution to the equations when \( \vec{X} \) is set to \( \vec{x} \) and \( \vec{Y} \) is set to \( \vec{y} \). The result now follows immediately. ■

With this background, I can prove Proposition 2.9.2.

Proof of Proposition 2.9.2: Suppose that \( \vec{X} = \vec{x} \) is a cause of \( \varphi \) in \((M, \vec{u})\) with witness \((\vec{W}, \vec{w}, \vec{x}')\) according to the original (resp., modified) HP definition. Let \( \vec{Z} = \mathcal{V} - \vec{W} \), let \( \vec{Z}' \) consist of the variables in \( \vec{Z} \) that lie on a causal path in \((M, \vec{u})\) from some variable in \( \vec{X} \) to some variable in \( \varphi \), and let \( \vec{W}' = \mathcal{V} - \vec{Z}' \). Notice that \( \vec{W}' \) is a superset of \( \vec{W} \). Let \( \vec{W}' - \vec{W} = \vec{Y} = \{Y_1, \ldots, Y_k\} \). The proof diverges slightly now depending on whether we consider the original or modified HP definition. In the case of the original HP definition, let \( y_j \) be the value of \( Y_j \) such that \( (M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}](Y_j = y_j) \); in the case of the modified HP definition, let \( y_j \) be the value of \( Y_j \) such that \( (M, \vec{u}) \models Y_j = y_j \). I claim that \((\vec{W} \cdot \vec{Y}, \vec{w} \cdot \vec{y}, \vec{x}')\) is a witness to \( \vec{X} = \vec{x} \) being a cause of \( \varphi \) in \((M, \vec{u})\) according to the original (resp., modified) HP definition. (Recall from the proof of Theorem 2.2.3 that \( \vec{x} \cdot \vec{y} \) denotes the result of concatenating the vectors \( \vec{x} \) and \( \vec{y} \).) Once I prove this, I will have proved the proposition.

Since \( \vec{X} = \vec{x} \) is a cause of \( \varphi \) by assumption, AC1 and AC3 hold; they are independent of the witness. This suffices to prove that AC2(a) and AC2(b°) (resp., AC2(a°)) hold for this witness.

Since \((\vec{W}, \vec{w}, \vec{x}')\) is a witness to \( \vec{X} = \vec{x} \) being a cause of \( \varphi \) in \((M, \vec{u})\) according to the original (resp., modified) HP definition, by AC2(a) (resp., AC2(a°)), it follows that

\[
(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}]\neg\varphi. \tag{2.4}
\]

Since no variable in \( \vec{Y} \) is on a causal path in \((M, \vec{u})\) from a variable in \( \vec{X} \) to a variable in \( \varphi \), for each \( Y \in \vec{Y} \), either there is no causal path in \((M, \vec{u})\) from a variable in \( \vec{X} \) to \( Y \) or there is no causal path in \((M, \vec{u})\) from \( Y \) to a variable in \( \varphi \) (or both). Without loss of generality, we can assume that the variables in \( \vec{Y} \) are ordered so that there is no causal path in \((M, \vec{u})\) from a variable in \( \vec{X} \) to any of the first \( j \) variables, \( Y_1, \ldots, Y_j \), and there is no causal path in \((M, \vec{u})\) from any of the last \( k - j \) variables, \( Y_{j+1}, \ldots, Y_k \), to a variable in \( \varphi \).

I first do the argument in the case of the modified HP definition. In this case, \( \vec{w} \) must be the value of the variables in \( \vec{W} \) in context \( \vec{u} \); by definition, \( \vec{y} \) is the value of the variables in \( \vec{Y} \). Thus, \( \vec{W} \cdot \vec{Y} \) is a legitimate witness for AC2(a°). Moreover, since \( (M, \vec{u}) \models \)
(\vec{W} = \vec{w}) \land (\vec{Y} = \vec{y})$, by Lemma 2.10.2, we have

$$(M, \vec{u}) \models [\vec{W} \leftarrow \vec{w}](Y_1 = y_1 \land \ldots \land Y_j = y_j).$$

Since there is no causal path in $(M, \vec{u})$ from $\vec{X}$ to any of $Y_1, \ldots, Y_j$, by Lemma 2.10.1, it follows that

$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}](Y_1 = y_1 \land \ldots \land Y_j = y_j). \quad (2.5)$$

Thus, by (2.4), (2.5), and Lemma 2.10.2, it follows that

$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}, Y_1 \leftarrow y_1, \ldots, Y_j \leftarrow y_j] \neg \varphi. \quad (2.6)$$

Finally, since there is no causal path in $(M, \vec{u})$ from any of $Y_{j+1}, \ldots, Y_k$ to any variable in $\varphi$, it follows from (2.6) and Lemma 2.10.1 that

$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}, \vec{Y} \leftarrow \vec{y}, \vec{X} \leftarrow \vec{x}'] \neg \varphi. \quad (2.9)$$

Thus, $\vec{X} = \vec{x}$ is a cause of $\varphi$ in $(M, \vec{u})$ according to the original HP definition with witness $(\vec{W} \cdot \vec{Y}, \vec{w} \cdot \vec{y}, \vec{x}')$. \(\blacksquare\)
Chapter 2. The HP Definition of Causality

Notes

The use of structural equations in models for causality goes back to the geneticist Sewall Wright [1921] (see Goldberger 1972 for a discussion). Herb Simon [1953], who won a Nobel Prize in Economics (and, in addition, made fundamental contributions to Computer Science), and the econometrician Trygve Haavelmo [1943] also made use of structural equations. The formalization used here is due to Judea Pearl, who has been instrumental in pushing for this approach to causality. Pearl [2000] provides a detailed discussion of structural equations, their history, and their use.

The assumption that the model is recursive is standard in the literature. The usual definition says that if the model is recursive, then there is a total order $\preceq$ of the variables such that $X$ affects $Y$ only if $X \preceq Y$. I have made two changes to the usual definition here. The first is to allow the order to be partial; the second is to allow it to depend on the context. As the discussion of the rock-throwing example (Example 2.3.3) shows, the second assumption is useful and makes the definition more general; a good model of the rock-throwing example might well include contexts where Billy hits before Suzy if they both throw. Although it seems more natural to me to assume a partial order as I have done here, since the partial order can be read off the causal network, assuming that the order is partial is actually equivalent to assuming that it is total. Every total order is a partial order, and it is well known that every partial order can be extended (typically, in many ways) to a total order. If $\preceq'$ is a total order that extends $\preceq$ and $X$ affects $Y$ only if $X \preceq Y$, then certainly $X$ affects $Y$ only if $X \preceq' Y$. Allowing the partial order to depend on the context $\vec{u}$ is also nonstandard, but a rather minimal change.

As I said in the notes to Chapter 1, the original HP definition was introduced by Halpern and Pearl in [Halpern and Pearl 2001]; it was updated in [Halpern and Pearl 2005a]; the modified definition was introduced in [Halpern 2015a]. These definitions were inspired by Pearl’s original notion of a causal beam [Pearl 1998]. (The definition of causal beam in [Pearl 2000, Chapter 10] is a modification of the original definition that takes into account concerns raised in an early version of [Halpern and Pearl 2001].) Interestingly, according to the causal beam definition, $A$ qualifies as an actual cause of $B$ only if something like $AC2(a^m)$ rather than $AC2(a)$ holds; otherwise it is called a contributory cause. The distinction between actual cause and contributory cause is lost in the original and updated HP definition. To some extent, it resurfaces in the modified HP definition; in some cases, what the causal beam definition would classify as a contributory cause but not an actual cause would be classified as part of a cause but not a cause according to the modified HP definition.

Although I focus on (variants of) the HP definition of causality in this book, it is not the only definition of causality that is given in terms of structural equations. Besides Pearl’s causal beam notion, other definitions were given by (among others) Glymour and Wimberly [2007], Hall [2007], Hitchcock [2001, 2007], and Woodward [2003]. Problems have been pointed out with all these definitions; see [Halpern 2015a] for a discussion of some of them. As I mentioned in Chapter 1, there are also definitions of causality that use counterfactuals without using structural equations. The best known is due to Lewis [1973a, 2000]. Paul and Hall [2013] cover a number of approaches that attempt to reduce causality to counterfactuals.
Proposition 2.2.2 and parts (a) and (b) of Theorem 2.2.3 are taken from [Halpern 2015a]. Part (d) of Theorem 2.2.3, the fact that with the original HP definition causes are always single conjuncts, was proved (independently) by Hopkins [2001] and Eiter and Lukasiewicz [2002].

Much of the material in Sections 2.1–2.3 is taken from [Halpern and Pearl 2005a]. The formal definition of causal model is taken from [Halpern 2000]. Note that in the literature (especially the statistics literature), what I call here “variables” are called random variables.

In the philosophy community, counterfactuals are typically defined in terms of “closest worlds” [Lewis 1973b; Stalnaker 1968]; a statement of the form “if A were the case then B would be true” is taken to be true if in the “closest world(s)” to the actual world where A is true, B is also true. The 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, can be viewed as the closest world to the actual world where $Y = y$.

The asymmetry embodied in the structural equations (i.e., the fact that variables on the left- and right-hand sides of the equality sign are treated differently) can be understood in terms of closest worlds. Suppose that in the actual world, the arsonist does not drop a match, there is no lightning, and the forest does not burn down. If either the match or lightning suffices to start the fire, then in the closest world to the actual world where the arsonist drops a lit match, the forest burns down. However, it is not necessarily the case that the arsonist drops a match in the world closest to the actual world where the forest burns down. The formal connection between causal models and closest-world semantics for counterfactuals is somewhat subtle; see [Briggs 2012; Galles and Pearl 1998; Halpern 2013; Zhang 2013] for further discussion.

The question of whether there is some circularity in using causal models, where the structural equations can arguably be viewed as encoding causal relationships, to provide a model of actual causality was discussed by Woodward [2003]. As he observed, if we intervene to set $X$ to $x$, the intervention can be viewed as causing $X$ to have value $x$. But we are not interested in defining a causal relationship between interventions and the values of variables intervened on. Rather, we are interested in defining a causal relation between the values of some variables and the values of otherwise; for example, we want to say that the fact that $X = x$ in the actual world is a cause of $Y = y$ in the actual world. As Woodward points out (and I agree), the definition of actual causality depends (in part) on the fact that intervening on $X$ by setting it to $x$ causes $X = x$, but there is no circularity in this dependence.

The causal networks that are used to describe the equations are similar in spirit to Bayesian networks, which have been widely used to represent and reason about (conditional) dependencies in probability distributions. Indeed, once we add probability to the picture as in Section 2.5, the connection is even closer. Historically, Judea Pearl [1988, 2000] introduced Bayesian networks for probabilistic reasoning and then applied them to causal reasoning. This similarity to Bayesian networks will be exploited in Chapter 5 to provide insight into when we can obtain compact representations of causal models.

Spirtes, Glymour, and Scheines [1993] study causality by working directly with graphs, not structural equations. They consider type causality rather than actual causality; their focus is on (algorithms for) discovering causal structure and causal influence. This can be viewed as work that will need to be done to construct the structural equations that I am taking as given.
here. The importance of choosing the “right” set of variables comes out clearly in the Spirtes et al. framework as well.

The difference between conditions and causes in the law is discussed by Katz [1987] and Mackie [1965], among others.

In the standard philosophical account, causality relates events, that is, for $A$ to be a cause of $B$, $A$ and $B$ have to be events. But what counts as an event is open to dispute. There are different theories of events. (Casati and Varzi [2014] provide a recent overview of this work; Paul and Hall [2013, pp. 58–60] discuss its relevance to theories of causality.) A major issue is whether something not happening counts as an event [Paul and Hall 2013, pp. 178–182]. (This issue is also related to that of whether omissions count as causes, discussed earlier.) The $A$s and $B$s that are the relata of causality in the HP definition are arguably closer to what philosophers have called true propositions (or facts) [Mellor 2004]. Elucidating the relationship between all these notions is well beyond the scope of this book.

Although I have handled the rock-throwing example here by just adding two additional variables ($BH$ and $SH$), as I said, this is only one of many ways to capture the fact that Suzy’s rock hit the bottle and Billy’s did not. The alternative of using variables indexed by time was considered in [Halpern and Pearl 2005a].

Examples 2.3.4 (double prevention), 2.3.5 (omissions as causes), and 2.4.1 (lack of transitivity) are due to Hall; their descriptions are taken from (an early version of) [Hall 2004]. Again, these problems are well known and were mentioned in the literature much earlier. See Chapter 3 for more discussion of causation by omission and the notes to that chapter for references.

Example 2.3.6 is due to Paul [2000]. In [Halpern and Pearl 2005a], it was modeled a little differently, using variables $LT$ (for train is on the left track) and $RT$ (for train is on the right track). Thus, in the actual context, where the engineer flips the switch and the train goes down the right track, we have $F = 1$, $LT = 0$, and $RT = 0$. With this choice of variables, all three variants of the HP definition declare $F = 1$ a cause of $A = 1$ (we can fix $RT$ at its actual value of 0 to get the counterfactual dependence of $A$ on $F$). This problem can be dealt with using normality considerations (see the last paragraph of Example 3.4.3). In any case, as Hitchcock and I pointed out [Halpern and Hitchcock 2010], this choice of variables is somewhat problematic because $LT$ and $RT$ are not independent. What would it mean to set both $LT$ and $RT$ to 1, for example? That the train is simultaneously going down both tracks? Using $LB$ and $FB$, as done in [Halpern and Hitchcock 2010], captures the essence of Hall’s story while avoiding this problem. The issue of being able to set variables independently is discussed in more detail in Section 4.6.

Example 2.3.8 is taken from [Halpern 2015a].

O’Connor [2012] says that, each year, an estimated 4,000 cases of “retained surgical items” are reported in the United States and discusses a lawsuit resulting from one such case.

Example 2.3.7 is due to Bas van Fraassen and was introduced by Schaffer [2000b] under the name trumping preemption. Schaffer (and Lewis [2000]) claimed that, because orders from higher-ranking officers trump those of lower-ranking officers, the captain is a cause of the charge and the sergeant is not. The case is not so clearcut to me; the analysis in the main text shows just how sensitive the determination of causality is to details of the model. Halpern and Hitchcock [2010] make the point that if $C$ and $S$ have range \{0, 1\} (and the only variables
in the causal model are $C$, $S$, and $P$), then there is no way to capture the fact that in the case of conflicting orders, the private obeys the captain. The model of the problem that captures trumping with the extra variable $SE$, described in Figure 2.7, was introduced in [Halpern and Pearl 2005a].

Example 2.4.2 is due to McDermott [1995], who also gives other examples of lack of transitivity, including Example 3.5.2, discussed in Chapter 3.

The question of the transitivity of causality has been the subject of much debate. As Paul and Hall [2013] say, “Causality seems to be transitive. If $C$ causes $D$ and $D$ causes $E$, then $C$ thereby causes $E$.” Indeed, Paul and Hall [2013, p. 215] suggest that “preserving transitivity is a basic desideratum for an adequate analysis of causation”. Lewis [1973a, 2000] imposes transitivity in his definition of causality by taking causality to be the transitive closure (“ancestral”, in his terminology) of a one-step causal dependence relation. But numerous examples have been presented that cast doubt on transitivity; Richard Scheines [personal communication, 2013] suggested the homeostasis example. Paul and Hall [2013] give a sequence of such counterexamples, concluding that “What’s needed is a more developed story, according to which the inference from ‘$C$ causes $D$’ and ‘$D$ causes $E$’ to ‘$C$ causes $E$’ is safe provided such-and-such conditions obtain—where these conditions can typically be assumed to obtain, except perhaps in odd cases . . . .” Hitchcock [2001] argues that the cases that create problems for transitivity fail to involve appropriate “causal routes” between the relevant events. Propositions 2.4.3, 2.4.4, and 2.4.6 can be viewed as providing other sets of conditions for when we can safely assume transitivity. These results are proved in [Halpern 2015b], from where much of the discussion in Section 2.4 is taken. Since these definitions apply only to but-for causes, which all counterfactual-based accounts (not just the HP accounts) agree should be causes, the results should hold for almost all definitions that are based on counterfactual dependence and structural equations.

As I said, there are a number of probabilistic theories of causation that identify causality with raising probability. For example, Eells and Sober [1983] take $C$ to be a cause of $E$ if $Pr(E \mid K_i \land C) > Pr(E \mid K_i \land \neg C)$ for all appropriate maximal specifications $K_i$ of causally relevant background factors, provided that these conditionals are defined (where they independently define what it means for a background factor to be causally relevant); see also [Eells 1991]. Interestingly, Eells and Sober [1983] also give sufficient conditions for causality to be transitive according to their definition. This definition does not involve counterfactuals (so the results of Section 2.4 do not apply).

There are also definitions of probabilistic causality in this spirit that do involve counterfactuals. For example, Lewis [1973a] defines a probabilistic version of his notion of dependency, by taking $A$ to causally depend on $B$ if and only if, had $B$ not occurred, the chance of $A$’s occurring would be much less than its actual chance; he then takes causality to be the transitive closure of this causal dependence relation. There are several other ways to define causality in this spirit; see Fitelson and Hitchcock [2011] for an overview. In any case, as Example 2.5.2 shows, no matter how this is made precise, defining causality this way is problematic. Perhaps the best-known example of the problem is due to Rosen and is presented by Suppes [1970]: A golfer lines up to driver her ball, but her swing is off and she slices it badly. The slice clearly decreases the probability of a hole-in-one. But, as it happens, the ball bounces off a tree trunk at just the right angle that, in fact, the golfer gets a hole-in-one. We want to say that the slice
caused the hole-in-one (and the definition given here would say that), although slicing the ball lowers the probability of a hole-in-one.

The difference between the probability of rain conditional on a low barometer reading and the probability that it rains as a result of intervening to set the barometer reading to low has been called by Pearl [2000] the difference between seeing and doing.

Examples 2.5.1, 2.5.2, and 2.5.3 are modifications of examples in [Paul and Hall 2013], with slight modifications; the original version of Example 2.5.3 is due to Frick [2009]. As I said, the idea of “pulling out the probability” is standard in computer science; see [Halpern and Tuttle 1993] for further discussion and references. Northcott [2010] defends the view that using deterministic models may capture important features of the psychology of causal judgment; he also provides a definition of probabilistic causality. Fenton-Glynn [2016] provides a definition of probabilistic causality in the spirit of the HP definitions; unfortunately, he does not consider how his approach fares on the examples in the spirit of those discussed here. Much of the material in Section 2.5 is taken from [Halpern 2014b].

Example 2.5.4 is in the spirit of an example considered by Hitchcock [2004]. In his example, both Juan and Jennifer push a source of polarized photons, which greatly increases the probability that a photon emitted by the source will be transmitted by a polarizer. In fact, the polarizer does transmit a photon. Hitchcock says, “It seems clearly wrongheaded to ask whether the transmission was really due to Juan’s push rather than Jennifer’s push. Rather . . . each push increased the probability of the effect (the transmission of the photon), and then it was simply a matter of chance. There are no causal facts of the matter extending beyond the probabilistic contribution made by each. This is a very simple case in which our intuitions are not clouded by tacit deterministic assumptions.”

Hitchcock’s example involves events at the microscopic level, where arguably we cannot pull out the probability. But for macroscopic events, where we can pull out the probability, as is done in Section 2.5, I believe that we can make sense out of the analogous questions, and deterministic assumptions are not at all unreasonable.

Balke and Pearl [1994] give a general discussion of how to evaluate probabilistic queries in a causal model where there is a probability on contexts.

Representation of uncertainty other than probability, including the use of sets of probability measures, are discussed in [Halpern 2003].

Lewis [1986b, Postscript C] discussed sensitive causation and emphasized that in long causal chains, causality was quite sensitive. Woodward [2006] goes into these issues in much more detail and makes the point that people’s causality judgments are clearly influenced by how sensitive the causality ascription is to changes in various other factors. He points out that one reason some people tend to view double prevention and absence of causes as not quite “first-class” causes might be that these are typically quite sensitive causes; in the language of Section 2.6, they are sufficient causes with only low probability.

Pearl and I [2005a] considered what we called strong causality, which is intended to capture some of the same intuitions behind sufficient causality. Specifically, we extended the updated definition by adding the following clause:

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

Thus, instead of requiring \([\bar{X} \leftarrow \bar{x}] \varphi\) to hold in all contexts, some of the same effect is achieved by requiring that \(\varphi\) holds no matter what values of \(\bar{W}\) are considered. The definition
given here seems somewhat closer to the intuitions and allows the consideration of probabilistic sufficient causality by putting a probability on contexts. This, in turn, makes it possible to bring out the connections between sufficient causality, normality, and blame.

Although Pearl [1999, 2000] does not define the notion of sufficient causality, he does talk about the probability of sufficiency, that is, the probability that $X = x$ is a sufficient cause of $Y = y$. Roughly speaking, this is the probability that setting $X$ to $x$ results in $Y = y$, conditional on $X \neq x$ and $Y \neq y$. Ignoring the conditioning, this is very close in spirit to the definition given here. Datta et al. [2015] considered a notion close to the notion of sufficient causality defined here and pointed out that it could be used to distinguish joint and independent causality. Honoré [2010] discusses the distinction between joint and independent causation in the context of the law. Gerstenberg et al. [2015] discuss the impact of sufficient causality on people’s causality judgments.

The definition of causality in nonrecursive causal models is taken from the appendix of [Halpern and Pearl 2005a]. Strotz and Wold [1960] discuss recursive and nonrecursive models as used in econometrics. They argue that our intuitive view of causality really makes sense only in recursive models, and that once time is taken into account, a nonrecursive model can typically be viewed as recursive.

Example 2.8.1, which motivated AC2(b), is due to Hopkins and Pearl [2003]. Example 2.8.2 is taken from [Halpern 2008]. Hall [2007] gives a definition of causality that he calls the $H$-account that requires (2.1) to hold (but only for $\vec{W}$, not for all subsets $\vec{W}' \subseteq \vec{W}$). Example 2.8.3, which is taken from [Halpern and Pearl 2005a], shows that the $H$-account would not declare $V_1 = 1$ a cause of $P = 1$, which seems problematic. This example also causes related problems for Pearl’s [1998, 2000] causal beam definition of causality.

Hall [2007], Hitchcock [2001], and Woodward [2003] are examples of accounts of causality that explicitly appeal to causal paths. In [Halpern and Pearl 2005a, Proposition A.2], it is claimed that all the variables in $\vec{Z}$ can be taken to lie on a causal path from a variable in $\vec{X}$ to $\varphi$. As Example 2.9.1 shows, this is not true in general. In fact, the argument used in the proof of [Halpern and Pearl 2005a, Proposition A.2] does show that with AC2(b), all the variables in $\vec{Z}$ can be taken to lie on a causal path from a variable in $\vec{X}$ to $\varphi$ (and is essentially the argument used in the proof of Proposition 2.9.2). The definition of causality was changed from using AC2(b) to AC2(b) in the course of writing [Halpern and Pearl 2005a]; Proposition A.2 was not updated accordingly.