In H. Geffner, R. Dechter, and J.Y. Halpern (Eds.), Probabilistic and Causal Inference: The Works of Judea Pearl, Association for Computing Machinery, 2021.

# 32 Pearl on Actual Causation

**Christopher Hitchcock** (California Institute of Technology)

## Abstract

This chapter surveys Judea Pearl's work on actual causation. After briefly introducing the concept of actual causation, it presents the structural equation framework used by Pearl to analyze actual causation. Earlier definitions of actual causation are presented to illustrate some of the difficulties involved in analyzing this concept. One of Pearl's definitions of actual causation is presented in detail, and its strengths and weaknesses are examined. The chapter concludes with reflections on Pearl's contributions to the topic.

## 32.1 Introduction

Judea Pearl offered three different but closely related definitions of *actual causation* using the formalism of *structural equation models*. The first appeared in chapter 10 of *Causality: Models, Reasoning, and Inference* [Pearl 2000, 2009]; the others appeared in a series of papers co-authored with Joseph Halpern [Halpern and Pearl 2001a, 2001b, 2005a, 2005b]. Pearl's definitions are based on the *but-for* definition of causation used in common law, and build on important earlier work by the philosopher David Lewis [Lewis 1973, 1986]. Pearl's definitions have been very influential and have inspired a number of further attempts to refine the definition within the same formalism; an incomplete selection includes Blanchard and Schaffer [2017], Beckers and Vennekins [2017, 2018], Fenton-Glynn [2017], Gallow [2021], Glymour and Wimberly [2007], Hall [2007], Halpern [2008, 2016], Halpern and Hitchcock [2015], Hitchcock [2001, 2007], Menzies [2004, 2017], and Woodward [2003, chapter 2]. This chapter will provide an introduction to the topic.

## 32.2 Actual Causation

We may illustrate the concept of actual causation with a traditional example. Billy and Suzy are throwing stones. Suzy throws her stone at the window, it hits the

window, and the window breaks. We would naturally summarize this episode in one of the following ways:

- Suzy's throw caused the window to break
- Suzy caused the window to break by throwing a stone at it
- The window broke because Suzy threw a stone at it

These statements describe a relation of actual causation between two events: Suzy throwing a stone and the window breaking. We may make the following generalizations about relations of actual causation:

- They relate particular events, rather than types or properties. In our example, it is a particular throw, of a particular stone, by a particular girl, at a particular time and place that causes a particular window to break at a particular time and place. The statements of actual causation listed above say nothing about the efficacy of throws or rocks in general, nor about the causes of broken windows in general.
- They depend on how events actually play out. Suzy *might* not have thrown, her throw *might* not have hit, Billy *might* have thrown the stone that broke the window; but as things actually happened, it was Suzy's throw that caused the window to break.
- Claims of actual causation are typically (but not always) made after the fact. Before Suzy throws, it may be hard to predict whether she will throw or whether her aim will be true. After the fact, it is relatively easy to judge that Suzy's throw caused the window to break.
- Relations of actual causation are particularly relevant to judgments of moral responsibility and legal liability. We would hold Suzy morally responsible for the broken window and require her parents to pay for its replacement (Suzy is still a minor).

This is not a rigorous or complete definition, but it provides some indication of the target of analysis.

## 32.3 Causal Models and But-for Causation

One of Pearl's many innovations was introducing the use of *structural equation models* (SEMs) to represent the causal structure of a situation such as the one described in the vignette about Billy and Suzy. SEMs have been widely used in a number of fields, including agronomy, econometrics, and epidemiology, and Pearl has a great deal to say about their use in these areas as well. I will not attempt to

provide a rigorous presentation of this formalism, but will introduce it by means of examples in the hopes of making it intuitive. These examples will serve three purposes. They will help us to introduce the formalism; they will provide test cases for theories of actual causation; and they will demonstrate some of the problems facing earlier accounts. While I will present two previous definitions of actual causation to set the stage for Pearl's definitions, my formulations of them will be anachronistic—they will be couched within the formalism developed later by Pearl.

All examples will involve Billy and Suzy throwing stones at a window, and we will make the following assumptions throughout: (1) whenever Billy or Suzy throws a rock, their aim is true and they throw with sufficient force to shatter the window; (2) the window does not break spontaneously, or due to any other cause not explicitly mentioned. We will represent various scenarios using *variables* with the following interpretations:

- $ST$—Suzy throws her rock
- $SF$—Suzy's rock flies through the air toward the window
- $SH$—Suzy's rock hits the window
- $BT$—Billy throws his rock
- $BH$—Billy's rock hits the window
- $BB$—Billy blocks Suzy's rock
- $WB$—the window breaks

Each variable takes the value 1 if the relevant event occurs, and 0 if it does not. We might think of these as propositions that can be true or false, rather than variables. But the variables in a SEM need not be binary—we could, for example, have a variable representing the velocity of Suzy's rock—but we will restrict ourselves to binary variables for simplicity. An assignment of a value to a variable corresponds to a particular event; for example, $ST = 1$ corresponds to Suzy's throwing her rock at a particular time and place. These will be the candidates for causes and effects.

**Example 32.1** Suzy throws her rock at the window, which breaks. (Billy has not yet arrived.)

In this little story, it would be natural to judge that Suzy's throw caused the window to break. We can model this very simple example as follows:

**Model 32.1** $M_{32.1}$

- $ST = 1$
- $WB = ST$

The first equation tells us that $ST = 1$, that is, that Suzy throws her rock. In this model, $ST$ is an *exogenous* variable; its value is determined by factors that are not explicitly modeled.[1] The second equation tells us how the value of $WB$ *depends* upon the value of $ST$. Specifically, it tells us that if and only if $ST = 1$ (Suzy throws), then $WB = 1$ (the window breaks). However, this equation is different from a normal logical biconditional in that it matters which variable we put on the left-hand side. The equations in a causal model are *structural* equations, meaning that they encode information about causal structure. This model is *acyclic*, meaning that the equations can be ordered so that each variable appears on the left-hand side of an equation before it appears on the right. Variables that are introduced earlier in this ordering will be said to be *upstream*, and those that appear later are *downstream*. In $M_{32.1}$, $ST$ is upstream of $WB$, and $WB$ is downstream of $ST$. We will only consider acyclic models in what follows. In an acyclic SEM, the values of the exogenous variables uniquely determine the values of all of the endogenous variables via the equations. (Probability can be added to the models, but we will skip this complication.) Thus, in $M_{32.1}$, $WB$ will take the value 1, which we can write $M_{32.1} \models WB = 1$. One basic criterion of adequacy for a causal model is that it entail values of the variables corresponding to events that actually occurred in the situation or story being modeled. (For this reason, we will sometimes refer to the values that variables take in a given model as the *actual* values of the variables in that model.)

If we want to know what *would* have happened if Suzy had *not* thrown, we remove the original equation for $ST$ and replace it with the imposed value $ST = 0$.

**Model 32.1.1** $M_{32.1.1} = M_{32.1}[ST \leftarrow 0]$

- ~~$ST = 1$~~, $ST = 0$
- $WB = ST$

The notation $M_{32.1}[ST \leftarrow 0]$ indicates that the new model is formed by starting with $M_{32.1}$, striking out the equation for $ST$, and replacing it with the setting $ST = 0$. Setting the value of a variable in this way is called an *intervention*. We can now compute from the resulting equations that $WB = 0$. We have thus verified the following *counterfactual*: If Suzy hadn't thrown her rock, the window would not have broken. The breaking of the window *counterfactually depends* upon Suzy's throw. A second basic

1. I am oversimplifying the treatment of exogenous variables. In Pearl's various formulations, exogenous variables do not represent factors that form part of the scenario. Thus the full model would treat $ST$ as an endogenous variable whose value is determined by one or more exogenous variables. Pearl then distinguishes between the model proper, and a specific setting of the exogenous variables. I am combining both of these together in what I am calling a model.

condition of adequacy for a causal model is that it entail only counterfactuals that are true in the situation or story being modeled.

This leads us to a first attempt to define actual causation:

**Definition 32.1** **But-for.**
If $X$ *and* $Y$ are distinct variables in the causal model $M$, then $X = x$ is an actual cause of $Y = y$ in $M$ just in case:

1. $M \models X = x, Y = y$
2. There exist values $x' \neq x$ of $X$ and $y' \neq y$ of $Y$ such that $M[X \leftarrow x'] \models Y = y'$

This is the *but-for* definition of causation that is frequently used in common law. It tells us that $X = x$ is a cause of $Y = y$ just in case (1) these are the actual values of these variables, and (2) if $X$ had taken some other value, $Y$ would not have been equal to $y$. To simplify the later exposition, let us say that $X = x$ is a *but-for* cause of $Y = y$ in model $M$ just in case Definition 32.1 rules that $X = x$ is an actual cause of $Y = y$ in model $M$. In Example 32.1, as modeled by $M_{32.1}$, if $ST$ had not been equal to 1, $WB$ would not have been equal to 1. In the language of common law, the window would not have broken *but for* Suzy's throw. In Example 32.1, Definition 32.1 gives the intuitively correct answer. We may also model the scenario described in Example 32.1 by interpolating variables between $ST$ and $WB$:

**Model 32.1.2** $M_{32.1.2}$

- $ST = 1$
- $SF = ST$
- $SH = SF$
- $WB = SH$

This model tells us that whether the window breaks counterfactually depends upon whether Suzy's stone hits it, which depends upon whether Suzy's rock is flying through the air, which depends upon whether she threw it.

It is helpful, but not strictly necessary, to represent the structure of a causal model with a directed graph. We draw an arrow from $X$ to $Y$ just in case $X$ appears on the right-hand side of the equation for $Y$. The graph for $M_{32.1.2}$ is shown in Figure 32.1.

Like $M_{32.1}$, $M_{32.1.2}$ also implies that if Suzy had not thrown, the window would not have shattered, as the reader can verify by replacing the first equation with


**Figure 32.1** Directed graph of $M_{32.1.2}$.

$ST = 0$. Suppose now that we want to evaluate the counterfactual situation where Suzy's rock does not hit the window. Following our procedure, we produce the new model:

**Model 32.1.3** $M_{32.1.3} = M_{32.1.2}[SH \leftarrow 0]$

- $ST = 1$
- $SF = ST$
- ~~$SH = SF$~~, $SH = 0$
- $WB = SH$

Note that we *replace* the equation for $SH$, rather than just plugging in the value 0 for $SH$ in the original equations. This reflects the idea that when we intervene to set $SH = 0$ we override the previously existing causal structure and impose the value 0 on $SH$. This is similar to Lewis's idea that we should think of the antecedent of a counterfactual being made true by a small miracle [Lewis 1979]. We represent this graphically by "breaking the arrow" into $SH$ (Figure 32.2).

When we evaluate the new system of equations, we get $WB = 0$ (the window wouldn't have broken), but $ST$ and $SF$ remain unchanged (Suzy still would have thrown, and her rock still would have flown through the air). What this example shows is that counterfactuals do not *backtrack* (in the terminology of Lewis [1979]). A hypothetical change introduced through an intervention may lead to changes in the values of *downstream* variables, but it will not lead to any changes in the values of *upstream* variables. The relation of counterfactual dependence is asymmetric (in acyclic models).

The asymmetry of counterfactual dependence is a good thing for Definition 32.1: it means that Definition 32.1 does *not* have the consequence that Suzy's rock hitting the window caused her to throw it. More generally, if $X = x$ is an actual cause of $Y = y$, then $Y = y$ will not be an actual cause of $X = x$. Thus Definition 32.1 can capture the intuitive idea that causation is an asymmetric relation.

Two further points about counterfactuals: First, we can readily extend our procedure for evaluating counterfactuals to cases where we intervene on multiple variables. We replace the equations for all of the variables on which we intervene.


**Figure 32.2** Directed graph of $M_{32.1.3} = M_{32.1.2}[SH \leftarrow 0]$.

Second, if we intervene to set one or more variables to their actual values in the model, all other variables will take their actual values.[2] That is:

**Fact 32.1** If $M \models \vec{X} = \vec{x}, \vec{Y} = \vec{y}$, then $M[\vec{X} \leftarrow \vec{x}] \models \vec{Y} = \vec{y}$.[3]

## 32.4 Pre-emption and Lewis

It has been known since at least 1925 that Definition 32.1 is inadequate [McLaughlin 1925]. In particular, it fails in cases of *pre-emption*. Here is an illustration:

**Example 32.2** Billy and Suzy are holding their stones, ready to throw. Billy decides to let Suzy throw first. Suzy throws her rock, which shatters the window. If Suzy hadn't thrown her rock, Billy would have thrown his rock at the window.

In this example, the window's breaking does not counterfactually depend upon Suzy's throw. If Suzy hadn't thrown, Billy's rock would have broken the window. Nonetheless, it is natural to judge that Suzy's throw caused the window to shatter. This is called a case of *pre-emption* because Suzy pre-empted Billy by throwing first.

Here is a simple and natural causal model for Example 32.2:

**Model 32.2** $M_{32.2}$

- $ST = 1$
- $BT = \neg ST$
- $WB = ST \vee BT$

The second equation tells us that Billy would throw just in case Suzy doesn't. The third equation says that the window would break just in case either Suzy or Billy throws. This model is pictured in Figure 32.3. Note that the arrow from $ST$ to $BT$ indicates that the first variable influences the second, but it does not tell us what the direction of influence is. That is, the arrow does not tell us whether the equation is $BT = \neg ST$ or $BT = ST$—whether Suzy's throw causes Billy's throw or prevents it. Thus, the equations of the model contain strictly more information than the corresponding graph. The graph does help us to see that $ST$ influences $WB$ via two different routes: one direct and one via $BT$.

2. Note, however, that not all propositions remain true in the new model that results from such an intervention. In particular, some counterfactuals may change in truth value. See, for example, Briggs [2012] for discussion.

3. I am using $\vec{X} = \vec{x}$ as a fairly intuitive shorthand. If $\vec{X} \equiv (X_1, \ldots, X_n)$ is an ordered set of variables, and $\vec{x} \equiv (x_1, \ldots, x_n)$ is an ordered set of values, then $\vec{X} = \vec{x}$ abbreviates the conjunction of propositions $X_i = x_i$ for $i = 1, \ldots, n$.

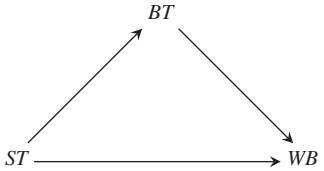


**Figure 32.3** Directed graph of $M_{32.2}$.

The reader can check that in $M_{32.2}$, $BT = 0$ (Billy doesn't throw) and $WB = 1$ (the window breaks). However, if Suzy hadn't thrown ($ST = 0$), then Billy would have thrown ($BT = 1$) and the window would have shattered anyway ($WB = 1$).

Lewis [1973] introduced a counterfactual theory of causation that improves upon the simple *but-for* definition. Lewis argued that causation is a *transitive* relation. If $X = x$ is an actual cause of $Y = y$ and $Y = y$ is an actual cause of $Z = z$, then $X = x$ should be an actual cause of $Z = z$. Definition 32.1 does not have this consequence since the relation of counterfactual dependence is not transitive (as we shall see in a moment). Lewis took counterfactual dependence to be *sufficient* for causation, but not necessary. $X = x$ can be an actual cause of $Z = z$ in the absence of counterfactual dependence if there is a suitable chain of counterfactual dependence.

**Definition 32.2** **Lewis**

If $X$ and $Z$ are distinct variables in the causal model $M$, then $X = x$ is an actual cause of $Z = z$ in $M$ just in case:

- There exists a sequence of variables $X \equiv Y_1, Y_2, \ldots, Y_{n-1}, Y_n \equiv Z$ such that: $Y_i = y_i$ is a but-for cause of $Y_{i+1} = y_{i+1}$ for all $i = 1, \ldots, n-1$.

Note that this entails that $M \models X = x, Z = z, Y_i = y_i$ for all $i$. *But-for* causation is a special case where $n = 2$.

Lewis's definition doesn't yield the intuitive result that $ST = 1$ is an actual cause of $WB = 1$ in $M_{32.2}$, but it does give this result in a slightly different model of Example 32.2, in which an additional variable is interpolated:

**Model 32.2.1** $M_{32.2.1}$

- $ST = 1$
- $BT = \neg ST$
- $SF = ST$
- $WB = SF \vee BT$

(See Figure 32.4.) In this model, $ST = 1$ is a *but-for* cause of $SF = 1$ (if Suzy hadn't thrown, her rock wouldn't have flown through the air); and $SF = 1$ is a *but-for*

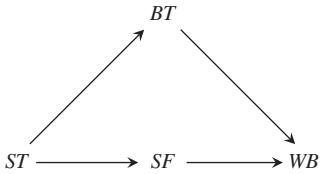


**Figure 32.4** Directed graph of $M_{32.2.1}$.

cause of $WB = 1$ (if Suzy's rock hadn't been flying through the air, the window wouldn't have broken). Thus we have a chain of counterfactual dependence, and Definition 32.2 rules that $ST = 1$ is an actual cause of $WB = 1$. The first step of this chain, from $ST$ to $SF$, is both intuitive, and easy to verify using model $M_{32.2.1}$. The second step, from $SF$ to $WB$, is less intuitive. We will first use the model to evaluate what happens under the counterfactual supposition that $SF = 0$:

**Model 32.2.2** $M_{32.2.2} = M_{32.2.1}[SF \leftarrow 0]$

- $ST = 1$
- $BT = \neg ST$
- ~~$SF = ST$~~, $SF = 0$
- $WB = SF \vee BT$

In this model, $ST = 1$ (Suzy still throws), $BT = 0$ (Billy doesn't throw), $SF = 0$ (Suzy's rock does not fly through the air), and $WB = 0$ (the window remains intact). Since counterfactuals do not backtrack, if Suzy's rock hadn't flown she still would have thrown, and Billy still would have refrained from throwing. We are to imagine that Suzy's rock vanishes or disintegrates after leaving her hand, or something intervenes to knock it out of the air. Since Billy's throw was conditioned on Suzy's throw, and not on the flight of her rock, he would not throw in this situation.

One question this raises is whether $M_{32.2}$ or $M_{32.2.1}$ is the "right" model of Example 32.2. Definition 32.2 yields a definition of actual causation that is *model-relative*. But the hypothetical examples that are used to assess the adequacy of definitions of causation are presented in natural language; they don't wear a preferred model on their sleeve. This raises several questions: What makes one causal model rather than another the "right" model of a particular situation? Is there a uniquely correct causal model? If not, what makes a causal model *apt* for analysis? Halpern and Hitchcock [2010] and Blanchard and Schaffer [2017] provide some preliminary discussion of these issues. Given an analysis of actual causation, when are the verdicts of that analysis stable under additions to and deletions from a causal model? Is this a desirable feature of an analysis? Can this kind of stability be used

to motivate a particular analysis? Halpern [2016, chapter 4] and Gallow [2021] take up these issues. As we will see, model-relativity will be a recurring issue.

## 32.5 Intransitivity and Overdetermination

Despite achieving some success with Example 32.2, Lewis's definition faces problems with other examples. The first such example raises questions about Lewis's hypothesis that actual causation is transitive.

**Example 32.3** *S*uzy throws her rock toward the window. Billy does not want the window to break, so he leaps into action and blocks Suzy's rock. The window remains intact.

We can model this example as follows:

**Model 32.3** $M_{32.3}$

- $ST = 1$
- $BB = ST$
- $WB = ST \wedge \neg BB$

(See Figure 32.5.) The last equation says that the window will break just in case Suzy throws and Billy doesn't block her rock.

In this model, $ST = 1$ is a *but-for* cause of $BB = 1$: if Suzy hadn't thrown, Billy wouldn't have blocked her rock. Moreover, $BB = 1$ is a *but-for* cause of $WB = 0$: if Billy hadn't blocked Suzy's rock, the window would have broken. (Remember that counterfactuals do no backtrack, so if Billy hadn't blocked the rock, Suzy still would have thrown). We can verify this second counterfactual by intervening to set $BB = 0$.

**Model 32.3.1** $M_{32.3.1} = M_{32.3}[BB \leftarrow 0]$

- $ST = 1$
- ~~$BB = ST$~~, $BB = 0$
- $WB = ST \wedge \neg BB$

We can compute that $WB = 1$ in this model. Since there is a chain of counterfactual dependence from $ST = 1$ to $BB = 1$ to $WB = 0$, Definition 32.2 rules that $ST = 1$ is

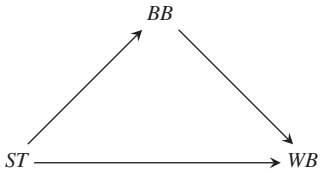


**Figure 32.5** Directed graph of $M_{32.3}$.

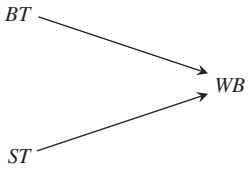


**Figure 32.6** Directed graph of $M_{32.4}$.

an actual cause of $WB = 0$.[4] But most people find this verdict unintuitive. Suzy's throw did not cause the window to remain intact (or prevent it from breaking). Lewis's definition gives the wrong answer. Moreover, this example is a counterexample to the transitivity of causation: Suzy's throw caused Billy to block her rock, and Billy's action caused the window to remain intact, but Suzy's throw did not cause the window to remain intact. This undermines one of the main motivations for moving from Definition 32.1 to Definition 32.2.

Lewis's definition also has trouble with causes of *symmetric overdetermination*:

**Example 32.4** Billy and Suzy both throw their rocks at the window. The rocks hit the window simultaneously, and the window breaks.

**Model 32.4** $M_{32.4}$

- $ST = 1$
- $BT = 1$
- $WB = ST \vee BT$

(See Figure 32.6.) The logical *or* in the last equation reflects the fact that either throw would be sufficient on its own to break the window.

$WB = 1$ does not counterfactually depend upon $ST = 1$: If Suzy hadn't thrown, the window still would have broken (because of Billy's throw). Nonetheless, most people judge that Susy's throw and Billy's throw are both causes of the window breaking.[5] I will leave it to the reader to verify that it does not help to interpolate variables such as *SF* or *SH* between *ST* and *WB*.

Here is another case of pre-emption that differs from Example 32.2[6]:

4. Interpolating a variable such as *SF* between *ST* and *WB* won't change this result.

5. Or perhaps they are parts of a joint cause. This is the verdict of one of the definitions of actual causation discussed in Halpern [2016].

6. This is an example of what Lewis [1986] calls *late pre-emption*; Example 32.2 is a case of *early pre-emption*. The nomenclature is not very intuitive. The key difference is that in early pre-emption the back-up process (Billy) is cut off before the effect (the window breaking) occurs; in late pre-emption the back-up process is still in progress when the effect occurs.

**Example 32.5** Suzy throws her rock slightly before Billy does. Her rock hits the window and smashes it. Billy's rock sails through the space where the window used to be.

Once again, it seems clear that Suzy's throw caused the window to break; but the window would have broken if Suzy hadn't thrown (due to Billy's rock). And once again, interpolating variables does not solve the problem. Unlike Example 32.4, however, there is an asymmetry between Suzy's throw and Billy's throw: Suzy's throw is a cause of the window breaking, but Billy's is not.

How should we model Example 32.5? $M_{32.4}$, which we used to model Example 32.4, is minimally adequate in the sense that it correctly describes the values of the variables, and that it also entails only true counterfactuals. However, if a definition of actual causation is going to yield a different verdict about Examples 32.4 and 32.5, then we will need to model these cases differently. In particular, it is apparent that $M_{32.4}$ is *symmetric* between $ST$ and $BT$. Any account of actual causation that rules that $ST = 1$ is an actual cause of $WB = 1$ in $M_{32.4}$ will also have to rule that $BT = 1$ is an actual cause. If we wish to rule that Susy's throw is a cause of the window breaking in Example 32.5 while Billy's throw is not, there will need to be a corresponding asymmetry in the causal model. A more adequate representation (from Halpern and Pearl [2001a]) would be:

**Model 32.5** $M_{32.5}$

- $ST = 1$
- $SH = ST$
- $BT = 1$
- $BH = BT \wedge \neg SH$
- $WB = SH \vee BH$

(See Figure 32.7.) In this model, we can derive that $SH = 1$ (Suzy's rock hits the bottle), while $BH = 0$ (Billy's rock does not hit the bottle). This is an important asymmetry between Suzy's throw and Billy's throw that we might hope to exploit.

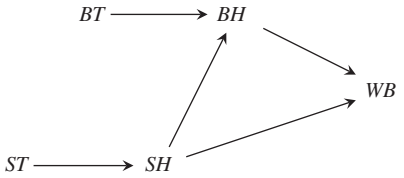


**Figure 32.7** Directed graph of $M_{32.5}$.

## 32.6 Pearl's Definitions of Actual Causation

Pearl has given three different definitions of Actual Causation in his published work, in chapter 10 of Pearl [2000, 2009][7]; and in a series of papers co-authored with Halpern [Halpern and Pearl 2001a, 2001b, 2005a, 2005b]. I will focus here on the definition from Halpern and Pearl [2001a].[8]

**Definition 32.3** **HP.**
If $X$ and $Y$ are distinct variables in causal model $M$, then $X = x$ is an actual cause of $Y = y$ in $M$ just in case:

1. $M \models X = x, Y = y$
2. There exists a partition $(\vec{Z}, \vec{W})$ of the variables in $M$, with $X \in \vec{Z}$, some setting $x'$ of $X$, and some setting $\vec{w}'$ of the variables in $\vec{W}$ such that
   (a) $M[X \leftarrow x', \vec{W} \leftarrow \vec{w}'] \models Y \neq y$
   (b) $M[X \leftarrow x, \vec{W} \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*] \models Y = y$ for all $\vec{Z}' \subseteq \vec{Z}$ (where $M \models \vec{Z}' = \vec{z}^*$).[9]

Condition 1 is straightforward: it just says that $x$ and $y$ are the values that $X$ and $Y$ actually take in the model. Condition 2 requires some unpacking.

The variables in the causal model are split into two sets, $\vec{W}$, and $\vec{Z}$. We may think of $\vec{Z}$ as making up the *causal process*. It will include $X$ and $Y$, and may also include some of the variables that lie on causal paths between $X$ and $Y$. The variables in $\vec{W}$ may be thought of as being *off to the side*. While they may lie on *some* causal

7. A predecessor of this definition appears in a technical report [Pearl 1998].

8. Halpern and Pearl [2001a] and the postscript to chapter 10 of Pearl [2009] describe the reasons for preferring the definition of Halpern and Pearl [2001a] to that of Pearl [2000]. Halpern and Pearl were moved to modify their definition in light of a putative counterexample described in Hopkins and Pearl [2003], giving rise to the new definition presented in Halpern and Pearl [2005a]. However, I think that the earlier definition of Halpern and Pearl [2001a] can handle this example by using a more sophisticated model. This closely parallels the move from modeling Example 32.5 using $M_{32.4}$ to using $M_{32.5}$. The Hopkins–Pearl case is an example of pre-emption, and its structure is not adequately captured without adding one additional variable.

9. I have simplified this definition in a couple of ways. Halpern and Pearl [2001a] allow the effect to be an arbitrary Boolean combination of propositions about the values of variables in the model. They don't require that the cause and effect involve distinct variables, although they note the possibility of adding such a restriction. They also allow the cause to be a conjunction of assignments of values to variables, but add a third clause to the definition that imposes a minimality condition on the cause. It turns out that this minimality condition implies that causes always involve single variables. (This is not the case with the definition of Halpern and Pearl [2005a], however.)

path between $X$ and $Y$, they are not part of the particular causal process that makes $X = x$ an actual cause of $Y = y$.[10]

Condition 2(a) says that $Y = y$ counterfactually depends upon $X = x$, not in the original model $M$ but in the new model that results when we also set $\vec{W}$ to $\vec{w}'$. The values $\vec{w}'$ may be the actual values of $\vec{W}$, but they need not be.

Condition 2(b) is a restriction on the permissible settings $\vec{W} = \vec{w}'$. The condition tells us that the setting of $\vec{W} = \vec{w}'$ cannot interfere with the causal process $\vec{Z}$ too much. Specifically, setting $\vec{W}$ to $\vec{w}'$ can't result in a different value of $Y$ when $X$ is set to its actual value, and when any members of $\vec{Z}$ are set to their actual value.

When $X = x$ and $Y = y$ satisfy the conditions of Definition 32.3 in model $M$, we will say that $X = x$ is an *HP cause* of $Y = y$ in $M$. We may note the following two facts about Definition 32.3:

**Fact 32.2** When $\vec{W} = \emptyset$, Definition 32.3 reduces to Definition 32.1.

Hence *but-for* causation is sufficient for *HP* causation.

**Fact 32.3** When $M \models \vec{W} = \vec{w}'$, the setting $\vec{W} = \vec{w}'$ satisfies condition 2(b).

Fact 32.3 follows from Fact 32.1.

Let us now see how the Halpern–Pearl definition of actual causation handles our various examples.

**Analysis of Example 32.1** Suzy throws her rock at the window, which breaks.

- $M_{32.1} : ST = 1, WB = ST$

We want to show that $ST = 1$ (Suzy's throw) is an actual cause of $WB = 1$ (the window breaking). Let $\vec{W} = \emptyset$. By Fact 32.2, Definition 32.3 now reduces to Definition 32.1. Since the but-for test rules that $ST = 1$ is an actual cause of $WB = 1$ in this simple example, the HP test does as well.

**Analysis of Example 32.2** Billy decides to let Suzy throw first. Suzy throws her rock, which shatters the window. If Suzy hadn't thrown her rock, Billy would have thrown.

- $M_{32.2} : ST = 1, BT = \neg ST, WB = ST \vee BT$

We want to show that $ST = 1$ is an actual cause of $WB = 1$. Let $\vec{W} = (BT)$, and $\vec{w}' = (0)$. Since $BT = 0$ in $M_{32.2}$, Fact 32.3 implies that condition 2(b) is satisfied. To check condition 2(a):

- $M_{32.2}[ST \leftarrow 0, BT \leftarrow 0] : \cancel{ST = 1}, ST = 0, \cancel{BT = \neg ST}, BT = 0, WB = ST \vee BT$

10. Although there may be more than one way of dividing variables into sets such that Definition 32.3 is satisfied. Variables that are off to the side in one partition may be part of the causal process in another.

We can compute that $WB = 0$. This computation validates the counterfactual: "If Suzy didn't throw, and Billy didn't throw, the window would not have broken." An equivalent counterfactual that more closely tracks the logic of the Definition 32.3 is: "Holding fixed that Billy didn't throw, if Suzy hadn't thrown, the window would not have broken."

We may think of the analysis in this way: $ST$ influences $WB$ via two different causal pathways—one direct and one via $BT$ (see Figure 32.3). By intervening to fix the value of $BT$ at 0, we block the influence of $ST$ on $WB$ via the indirect path. When we "wiggle" $ST$, we prevent $BT$ from "wiggling" with it. We thus isolate the influence of $ST$ on $WB$ along the direct path. It is in virtue of this influence that $ST = 1$ is an actual cause of $WB = 1$.

**Analysis of Example 32.4** Billy and Suzy both throw their rocks at the window. The rocks hit the window simultaneously, and the window breaks.

- $M_{32.4} : ST = 1, BT = 1, WB = ST \vee BT$

We want to show that $ST = 1$ is an actual cause of $WB = 1$. Let $\vec{W} = (BT)$, and $\vec{w}' = (0)$. Since this is not the actual value of $BT$, we cannot rely on Fact 32.3 to guarantee that condition 2(b) is met. To check condition 2(b), we must set $ST = 1$ and $BT = 0$; and we must check that $WB = 1$ both when we set $WB$ to 1, and when we leave $WB$ alone. Obviously, if we set $WB$ to 1, we will have $WB = 1$. So let us check the other case:

- $M_{32.4}[ST \leftarrow 1, BT \leftarrow 0] : \cancel{ST = 1}, ST = 1, \cancel{BT = 1}, BT = 0, WB = ST \vee BT$

We can compute that $WB = 1$, so condition 2(b) is met.

Let us now check condition 2(a).

- $M_{32.4}[ST \leftarrow 0, BT \leftarrow 0] : \cancel{ST = 1}, ST = 0, \cancel{BT = 1}, BT = 0, WB = ST \vee BT$

We can compute that $WB = 0$, so condition 2(a) is met.

Although the window's breaking does not counterfactually depend upon Suzy's throw in the actual situation, it does depend on her throw in the closely related situation where Billy does not throw. Changing whether Billy throws does not interfere sufficiently with the process connecting Suzy's throw to the shattered window, so this is a legitimate situation in which to check for actual causation.

**Analysis of Example 32.5** Suzy throws her rock slightly before Billy does. Her rock hits the window and smashes it.

We will analyze this example using the more sophisticated model $M_{32.5}$ (Figure 32.7).

- $M_{32.5} : ST = 1, SH = ST, BT = 1, BH = BT \wedge \neg SH, WB = SH \vee BH$

We first want to show that $ST = 1$ is a cause of $WB = 1$. We may choose $\vec{W} = (BH)$ with the setting $\vec{w}' = (0)$.[11] Since this is the actual value of $BT$, Fact 32.3 implies that condition 2(b) is satisfied. To check 2(a):

- $M_{32.5}[ST \leftarrow 0, BH \leftarrow 0]$ :

$$\cancel{ST = 1}, ST = 0, SH = ST, BT = 1, \cancel{BH = BT \wedge \neg SH}, BH = 0, WB = SH \vee BH$$

This implies $WB = 0$, so 2(a) is satisfied.[12] The analysis is similar to that in Example 32.2. By holding $BH$ fixed at 0, we isolate the influence of Susy's throw along the path from $ST$ to $SH$ to $WB$.

We would also like to show that $BT = 1$ is not an actual cause of $WB = 1$. We will not go through all of the possible combinations, but let us see why the parallel strategy of choosing $\vec{W} = (SH)$ will not work. First, we could try the actual setting $SH = 1$. With this setting, condition 2(a) fails:

- $M_{32.5}[BT \leftarrow 0, SH \leftarrow 1]$ :

$$ST = 1, \cancel{SH = ST}, SH = 1, \cancel{BT = 1}, BT = 0, BH = BT \wedge \neg SH, WB = SH \vee BH$$

This model implies that $WB = 1$. When we fix $SH$ at 1, $WB$ does not counterfactually depend upon $BT$. So let us try instead the setting $SH = 0$. Since this is not the actual setting of $SH$, we will need to check whether this setting satisfies condition 2(b). We can show that it does not by choosing $\vec{Z}' = (BH)$. Since $BH$ takes the value 0 in the actual model, we need to check:

- $M_{32.5}[BT \leftarrow 0, SH \leftarrow 0, BH \leftarrow 0]$ :

$$ST = 1, \cancel{SH = ST}, SH = 0, \cancel{BT = 1}, BT = 0,$$
$$\cancel{BH = BT \wedge \neg SH}, BH = 0, WB = SH \vee BH$$

In this model, $WB = 0$, violating 2(b). Setting $SH = 0$ is too big a change to the model. Thus, no setting for $SH$ works.

The mathematically astute reader will notice that I have skipped Example 32.3. While Definition 32.3 yields the intuitively correct result when we use $M_{32.3}$, it yields the wrong result if we interpolate a variable.

11. There are other choices that will work: $BT = 0, BT = 1 \wedge BH = 0$, and $BT = 0 \wedge BH = 0$.

12. See Hall [2007] for criticism of this analysis of Example 32.5.

**Analysis of Example 32.3** Suzy throws her rock toward the window. Billy does not want the window to break, so he blocks Suzy's rock. The window remains intact.

We will model this example as follows:

**Model 32.3.2** $M_{32.3.2}$

- $ST = 1$
- $SF = ST$
- $BB = ST$
- $WB = SF \wedge \neg BB$

See Figure 32.8. Although the intuitive verdict is that Suzy's throw did not cause the window to remain intact, Definition 32.3 rules that $ST = 1$ is an actual cause of $WB = 1$. To see this, choose $\vec{W} = (SF)$ and $\vec{w}' = (1)$. Since this is the actual value of $SF$, Fact 32.3 implies that condition 2(b) is met. Checking 2(a):

- $M_{32.3.2}[ST \leftarrow 0, SF \leftarrow 1] : \cancel{ST = 1}, ST = 0, \cancel{SF = ST}, SF = 1, BB = ST, WB = SF \wedge \neg BB$

In this model, $WB = 1$. Since Suzy didn't throw, Billy didn't block. But Suzy's stone was flying through the air (at a point too late for Billy to block it) so the window broke.

Halpern and Pearl [2005a] address this problem by allowing causal models to include restrictions on interventions. That is, in addition to the structural equations, a causal model will also specify that certain combinations of values of variables are impermissible, and cannot be realized by interventions. For example, model $M_{32.3.2}$ might specify that one cannot simultaneously set $ST = 0$ and $SF = 1$ by intervention. Hitchcock [2001] notes that the counterfactual involved in this case is psychologically unnatural: We are to imagine that Suzy does not throw, lulling Billy into complacency; then somehow Suzy's rock appears mid-air flying toward the window, too late for Billy to block it. Hall [2007], Halpern [2008], Hitchcock [2007], Halpern and Hitchcock [2015], and Menzies [2017] try to resolve this kind of problem by appeal to considerations of *normality*: only combinations of settings

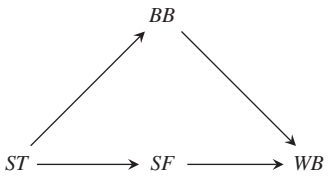


**Figure 32.8** Directed graph of $M_{32.3.2}$.

that correspond to normal states can underwrite relations of actual causation. All of these approaches imply that actual causation depends on more than just the objective content of causal models.

This example also highlights the recurring problem of model-relativity.

## 32.7 Pearl's Achievement

We have highlighted a few of Pearl's accomplishments on the topic of actual causation. He has introduced the formalism of SEMs to the project of defining actual causation. And he has offered new definitions that have improved upon previous definitions and have inspired further developments by others. But none of Pearl's definitions perfectly capture judgments of actual causation, and—spoiler alert—neither do any of the definitions that have followed. So where does this leave us?

The situation is familiar in philosophy. In Plato's famous dialogues, Socrates asks his students: What is justice? What is piety? What is knowledge? His students propose definitions, and Socrates presents clever counterexamples to shoot them down. Two and a half millennia later, we are still shooting them down. This is not to say that we have not learned a great deal in the process, but philosophy has not converged on accepted definitions of any of these concepts.

The situation is no less frustrating for being familiar. And it seems particularly frustrating in the case of causation. We might suspect that a concept like *justice* is multi-faceted, and perhaps at least partly subjective; for this reason it might defy precise definition. But surely *causation* is not like this? Aren't causal relations part of the objective structure of the world? Don't we have well-defined empirical procedures, such as randomized controlled trials, for establishing causal claims?

Perhaps Pearl's most important contribution to our understanding of actual causation is indirect. Through his work, we better understand the place of actual causation in our conceptual economy. By setting his definitions of actual causation in the much broader context of causal modeling and causal inference, Pearl has shown us that *actual causation* is in fact a very specialized causal concept. The very fact that Pearl's first definition appears in the tenth and last chapter of Pearl [2000] tells us that there is a great deal one can say about causation without settling on a definition of actual causation.

This fact is hidden in our language. We say: "Suzy's throw *caused* the window to break." The verb suggests a fully general notion of causation: nothing indicates that a specialized causal notion—actual causation—is being invoked.

Moreover, Pearl's work helps us to see that actual causation is not just causation among particular events (as a number of philosophers have suggested). As we have seen without examples, we can construct causal models of particular situations

that capture aspects of their causal structure. These models do not, by themselves, tell us what the *actual causes* are. For example, one cannot simply inspect $M_{32.5}$ and read off that Suzy's throw is an actual cause of the window shattering. To make this judgment, we further need a definition of actual causation in terms of the underlying causal structure. But even without such a definition, we can use our causal models to evaluate counterfactuals and predict the effects of interventions. This tells us that there is causal structure among individual events that is *not* actual causation.

Once we recognize that actual causation is a specialized causal concept that exists as a kind of overlay on a more basic causal skeleton of causal structure, it becomes more palatable to admit that actual causation may be like justice: multi-faceted, partly subjective, impossible to define precisely. We may admit this without denying that there is objective causal structure in the world, the kind of structure that can be rigorously investigated by using formal methods and empirical investigation. This does not mean that attempts to define actual causation are pointless.[13] For example, by embedding a concept of actual causation in a richer framework for investigating causation, we are better placed to ask and answer the question of why we have and use a notion of actual causation.[14] But thanks to Pearl, we may be a bit more forgiving on ourselves if our definitions of actual causation come up short. Our understanding of causation in general does not hang in the balance.

## References

S. Beckers and J. Vennekins. 2017. The transitivity and asymmetry of actual causation. *Ergo* 4, 1, 17. DOI: https://doi.org/10.3998/ergo.12405314.0004.001.

S. Beckers and J. Vennekins. 2018. A principled approach to defining actual causation. *Synthese* 195, 2, 835–862. DOI: https://doi.org/10.1007/s11229-016-1247-1.

T. Blanchard and J. Schaffer. 2017. Cause without default. In *Making a Difference: Essays on the Philosophy of Causation*. Oxford University Press, Oxford, 175–214. DOI: https://doi.org/10.1093/oso/9780198746911.001.0001.

R. Briggs. 2012. Interventionist counterfactuals. *Philos. Stud.* 160, 139–166. DOI: https://doi.org/10.1007/s11098-012-9908-5.

L. Fenton-Glynn. 2017. A proposed probabilistic extension of the Halpern and Pearl definition of "actual cause." *Br. J. Philos. Sci.* 68, 4, 1061–1124. DOI: https://doi.org/10.1093/bjps/axv056.

D. Gallow. 2021. A Model-invariant Theory of Causation. *Philosl. Rev.* 130, 45–96. DOI: https://doi.org/10.1215/00318108-8699682.

13. See Glymour et al. [2010] for skepticism on this score.

14. See Hitchcock [2017] for one attempt to do this.

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

C. Glymour, D. Danks, B. Glymour, F. Eberhardt, J. Ramsey, R. Scheines, P. Spirtes, C. M. Teng, and J. Zhang. 2010. Actual causation: A stone soup essay. *Synthese* 175, 169–192. DOI: https://doi.org/10.1007/s11229-009-9497-9.

N. Hall. 2007. Structural equations and causation. *Philos. Stud.* 132, 109–136. DOI: https://doi.org/10.1007/s11098-008-9216-2.

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

J. Y. Halpern. 2016. *Actual Causality*. M.I.T. Press, Cambridge, MA. DOI: https://doi.org/10.7551/mitpress/10809.001.0001.

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

J. Y. Halpern and C. Hitchcock. 2015. Graded causation and defaults. *Br. J. Philos. Sci.* 66, 413–457. DOI: https://doi.org/10.1093/bjps/axt050.

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

J. Y. Halpern and J. Pearl. 2001b. Causes and explanations: A structural-model approach. Part II: Explanation. In *Proceedings Seventeenth International Joint Conference on Artificial Intelligence (IJCAI '01).* 27–34.

J. Y. Halpern and J. Pearl. 2005a. Causes and explanations: A structural-model approach. Part I: Causes. *Br. J. Philos. Sci.* 56, 4, 843–887. DOI: https://doi.org/10.1093/bjps/axi147.

J. Y. Halpern and J. Pearl. 2005b. Causes and explanations: A structural-model approach. Part II: Explanations. *Br. J. Philos. Sci.* 56, 4, 889–911. DOI: https://doi.org/10.1093/bjps/axi148.

C. Hitchcock. 2001. The intransitivity of causation revealed in equations and graphs. *J. Philos.* 98, 6, 273–299. DOI: https://doi.org/10.2307/2678432.

C. Hitchcock. 2007. Prevention, preemption, and the principle of sufficient reason. *Philosl. Rev.* 116, 495–532. DOI: https://doi.org/10.1215/00318108-2007-012.

C. Hitchcock. 2017. Actual causation: What's the use? In H. Beebee, C. Hitchcock, and H. Price (Eds.), *Making a Difference: Essays on the Philosophy of Causation*. Oxford University Press, Oxford, 116–131. DOI: https://doi.org/10.1093/oso/9780198746911.001.0001.

M. Hopkins and J. Pearl. 2003. Clarifying the usage of structural models for commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning: Papers from the AAAI Spring Symposium*. AAAI Press, Menlo Park, CA, 83–89.

D. Lewis. 1973. Causation. *J. Philos.* 70, 556–567. DOI: https://doi.org/10.2307/2025310.

D. Lewis. 1979. Counterfactual dependence and time's arrow. *Noûs* 13, 455–476. DOI: https://doi.org/10.2307/2215339.

D. Lewis. 1986. Causation. In *Philosophical Papers*, Vol. II. Oxford University Press, Oxford. Includes Postscripts A-E to "Causation," 159–213. DOI: https://doi.org/10.1093/0195036468.001.0001.

J. A. McLaughlin. 1925. Proximate cause. *Harv. L. Rev.* 39, 149–199. DOI: https://doi.org/10.2307/1328484.

P. Menzies. 2004. Causal models, token causation, and processes. *Philos. Sci.* 71, 820–832. DOI: https://doi.org/10.1086/425057.

P. Menzies. 2017. The problem of counterfactual isomorphs. In H. Beebee, C. Hitchcock, and H. Price (Eds.), *Making a Difference: Essays on the Philosophy of Causation*. Oxford University Press, Oxford, 153–172. DOI: https://doi.org/10.1093/oso/9780198746911.001.0001.

J. Pearl. 1998. *On the Definition of Actual Cause*. Technical Report R-259, Department of Computer Science, University of California, Los Angeles, CA.

J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge. DOI: https://doi.org/10.10170S0266466603004109.

J. Pearl. 2009. *Causality: Models, Reasoning, and Inference* (2nd. ed). Cambridge University Press, Cambridge. DOI: https://doi.org/10.1017/CBO9780511803161.

J. Woodward. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford. DOI: https://doi.org/10.1093/0195155270.001.0001.