

How to Trace a Causal Process

J. DMITRI GALLOW

Abstract: According to the theory developed here, we may trace out the processes emanating from a cause in such a way that any consequence lying along one of these processes counts as an effect of the cause. This theory gives intuitive verdicts in a diverse range of problem cases from the literature. Its claims about causation will never be retracted when we include additional variables in our model. And it validates some plausible principles about causation, including Sartorio's 'Causes as Difference Makers' principle and Hitchcock's 'Principle of Sufficient Reason'.

1 Introduction

The morning of the space shuttle Challenger's launch was uncommonly cold. The near-freezing weather led to two O-rings in the shuttle's Solid Rocket Boosters (SRBs) being less elastic than they would otherwise have been. These less elastic O-rings allowed gas to leak from the SRBs shortly after launch. This leaked gas burnt a hole in shuttle's external fuel tank. And the breach of the fuel tank led to an explosion which destroyed the Challenger.

We are able to trace out a process: from the unusually cold weather to the rigidity of the O-rings, to the gas leak, to the breach of the fuel tank, to the explosion, to the shuttle's destruction. Having traced this process, we conclude that the cold weather was a cause of the shuttle's destruction. (Of course, it was not the *only* cause; there were many causes of this tragedy, but the weather is among them.) This case is typical. Often, having traced out a sequence of *c*'s consequences, we count all of the consequences in this sequence among *c*'s effects.

Often, but not always. The Soviet Union deployed missiles to Cuba. In response, the United States planned an invasion of Cuba. When Khrushchev

Word Count: 8,423
Draft of January 4, 2022

learnt of these plans, he initiated secret negotiations with the U.S. In these negotiations, the Soviet Union agreed to remove its missiles from Cuba in exchange for the U.S. removing its missiles from Turkey. This deal averted war between the U.S. and the U.S.S.R. So we are able to trace out a process: from the deployment of missiles to Cuba, to the planned invasion, to the negotiations, to the deal, to the peace. But in this case, having traced out this sequence of consequences, we are not inclined to count the deployment of missiles to Cuba among the causes of the peace. We are inclined to say that peace was maintained *in spite of* the missiles in Cuba, not *because of* them. (You may suspect that, for some complicated reason, we have the Cuban missile crisis to thank for keeping the Cold War cold. Even so, you should agree that the process we traced out above is not enough, on its own, to show that the Cuban missile crisis prevented war.)

Here, I will develop a theory according to which causation is closely related to the tracing of causal processes like these. However, the rules for causal process tracing are slightly more complicated than we may have naïvely thought. Not just any sequence of consequences counts as a *causal* process. On this theory, the rules for process tracing depend upon a prior distinction between states of the world which are regarded as the *default*, and events which are regarded as *departures* from that default. Just to have a term, call the states or events which deviate from the default *deviant*. On the view I'll propose, the effects of causes are determined by tracing out their consequences according to certain rules. I'll give a precise statement of these rules below, but just to give you some preliminary orientation: there are two ways of tracing out a causal process. On the one hand, you may trace out all and only the *deviant* consequences. On the other hand, you may trace out *all* potential consequences, including the non-deviant ones. A process traced out in either of these ways counts as a *causal* process. And everything along a causal process counts as an effect of the causes which initiated it.

I'll begin in section 2 below by introducing the relation of *influence*. Influence provides the pathways along which causal processes propagate. Then, in section 3, I will introduce the distinction between default and deviant variable values and explain why I've been persuaded that a theory of causation must incorporate something like this distinction. In section 4, I will explain what's involved in 'tracing out' a process and provide rules for how to trace a causal process. Section 5 applies the theory developed in section 4 to some illustrative cases and explores connections with other recent work on causation.

2 Influence

As I will understand them, causal processes propagate along networks of causal influence. As I use the term, *influence* is a kind of causal relation which holds between variables. (I distinguish influence from causation, which is a relation which holds between variable *values*.) In English, variables are named with expressions like “whether I go to the cotillion”, “when the bus arrives”, and “how much the lizard weighs”. When variables influence each other, this is naturally expressed in English with the verb ‘influences’ or ‘affects’—as in “whether James goes to the cotillion affects whether I do” or “how much the lizard eats influences how much it weighs”. For example, consider the following vignette:

Preemptive Overdetermination

Both the United States and the United Kingdom dispatch a covert agent to assassinate a foreign president. The CIA agent observes the MI6 agent providing poison to the president’s chef, and so, to protect their cover, they abandon their own assassination plot, which involved explosives hidden in the presidential palace. The chef puts the poison in the president’s food. The president eats the food and dies.

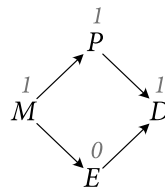
On a natural understanding of this case, whether the MI6 agent provides poison to the chef influences whether the chef puts the poison in the president’s food and whether the CIA agent ignites their explosives. And whether the president dies is influenced both by whether the chef puts the poison in the food and by whether the explosives are ignited. We can use ‘*D*’ to name the variable *whether the president dies*, ‘*P*’ for the variable *whether the food is poisoned*, ‘*E*’ for *whether the explosives are ignited*, and ‘*M*’ for *whether the MI6 agent provides poison to the chef*. Each of these variables has two potential values, 1 and 0. $D = 1$ if the president dies while $D = 0$ if they do not. $P = 1$ if the food is poisoned, while $P = 0$ if it is not poisoned. $E = 1$ if the explosives are ignited, while $E = 0$ if they are not. And $M = 1$ if the MI6 agent provides the poison, and $M = 0$ if they do not. If we make the simplifying assumption that these relations of influence are deterministic, then we can formally model them with a system of structural equations like the following:

$$D := P \vee E$$

$$P := M$$

$$E := \neg M$$

$$M = 1$$



In these structural equations, \vee and \neg are Boolean disjunction and negation, respectively. ($P \vee E = \max\{P, E\}$ and $\neg M = 1 - M$.) The structural equations tell us that M influences P , and that the value $M = 1$ causally determines that $P = 1$, whereas $M = 0$ causally determines that $P = 0$. Likewise, M causally influences E , with $M = 1$ causally determining that $E = 0$, and $M = 0$ causally determining that $E = 1$. And P and E together causally influence D , with either $P = 1$ or $E = 1$ causally determining that $D = 1$. Because causal influence isn't symmetric, structural equations are not symmetric, either. While we could rearrange a normal equation $E = \neg M$ to get $M = \neg E$, we cannot re-arrange a *structural* equation. That's why I've used ' $:=$ ' instead of the symmetric ' $=$ ' for the structural equations. The final equation, $M = 1$, is not structural. It tells us that M 's value is 1. This, together with the structural equations, allows us to work out the values of all of the other variables in the system. We will always be able to do this so long as there are not any cycles of influence. I'll ignore issues having to do with cyclic systems of equations here.

I've illustrated the network of causal influence described by these equations with a graph. This graph consists of four directed edges: $M \rightarrow P$, $M \rightarrow E$, $P \rightarrow D$ and $E \rightarrow D$. In general, we can build a causal graph by including a directed edge between two variables U and V , $U \rightarrow V$, iff U shows up on the right-hand-side of V 's structural equation. (A variable V 's structural equation is just the equation which has V on the left-hand-side.) In that case, I'll say that U is one of V 's *parents*.¹ Of course, whether one variable is a parent of another is relative to a particular system of structural equations. In some other, no less accurate, system of structural equations for *Preemptive Overdetermination*, there could be variables intermediate between M and P . If there's a sequence of directed edges leading from V to D , $V \rightarrow U_1 \rightarrow U_2 \rightarrow \dots \rightarrow U_N \rightarrow D$, then I'll call D a *descendant* of V 's.

The directed edges between variables in these graphs represent relations of influence between variables. And they provide the pathways along which causal processes propagate. We will only be able to trace a causal process leading from one variable value, $C = c$, to another, $E = e$, if there is a directed path

1. More carefully: we include the directed edge $U \rightarrow V$ iff, according to V 's equation, V 's value is a function of U 's value. So long as we write out the structural equations sensibly, these two characterisations come to the same thing. But suppose that, perversely, we write out V 's equation as $V := U + (T - T)$. In that case, even though T 'appears on the right-hand-side' of V 's equation, T 's value makes no difference to V 's value. In this case, we should not include a directed edge from T to V .

of influence leading from the variable C to the variable E .²

3 Defaults and Deviations

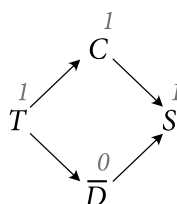
Unfortunately, in order to know whether $C = c$ is a cause of $E = e$, we will need to know more than a structural equations model on its own can tell us. Causation is underdetermined by the relations of influence between variables and the values of those variables.³ To appreciate this, consider the following vignette:

Tornado

A tornado approaches the farm. Seeing it coming, Aunt Em runs to the storm cellar. The tornado destroys the house, but the cellar protects Aunt Em, and she survives unscathed.⁴

On a natural understanding of this case, whether there is a tornado influences whether Aunt Em is in the cellar and whether the house is destroyed. And whether Aunt Em survives is influenced by whether she's in the cellar and whether the house is destroyed. Let me use ' T ', ' C ', and ' S ' for whether there's a tornado, whether Aunt Em is in the cellar, and whether Aunt Em survives, respectively. And I'll use ' \bar{D} ' for whether the house *is not* destroyed. All of these variables are binary, and all take on the truth-value of ϕ in their associated 'whether ϕ ' expression. Thus, $\bar{D} = 1$ if the house is *not* destroyed, and $\bar{D} = 0$ if it *is* destroyed. Then, making the simplifying assumption that the relations of influence are all deterministic, we can write down the following structural equations model:

$$\begin{aligned} S &:= C \vee \bar{D} \\ C &:= T \\ \bar{D} &:= \neg T \\ T &= 1 \end{aligned}$$



It does not appear that the tornado ($T = 1$) caused Aunt Em to survive ($S = 1$). We are inclined to say that Aunt Em survived *in spite of* the tornado,

2. In Gallow (2016), I provide a theory of influence. See also Gallow (ms).
 3. See Hall (2007) and Hiddleston (2005b).
 4. This case is modelled on *Boulder* from Hitchcock (2001), who attributes the case to an early draft of Hall (2004).

and not *because of* it. (*Tornado* is similar to the case of the Cuban missile crisis discussed in the introduction.) But notice that this system of structural equations is isomorphic to the system of structural equations for *Preemptive Overdetermination*. We may associate the variable T with M , C with P , \bar{D} with E , and S with D . When we do so, the ‘corresponding’ variables are related by exactly the same equations and take on exactly the same values. But, while $M = 1$ is a cause of $D = 1$ (the MI6 agent’s providing poison to the chef is a cause of the president’s death), $T = 1$ is *not* a cause of $S = 1$ (the tornado is not a cause of Aunt Em’s survival).

It’s important to recognise that the isomorphism between *Tornado* and *Preemptive Overdetermination* doesn’t depend upon us using the slightly unnatural variable \bar{D} . This variable makes it easier for us to *recognise* the isomorphism, but it would be there even if we used a variable which took on the value 1 if the house is destroyed, and took on the value 0 if the house is not destroyed.⁵ Any theory of causation formulated in terms of a structural equations model alone will not distinguish between models which are isomorphic in this way. So, if we only look at systems of structural equations, we’ll either have to say that the president’s death was not an effect of the MI6 agent’s providing poison to the chef, or else we’ll have to say that Aunt Em’s survival was an effect of the tornado.⁶

5. What do I mean when I say that there’s ‘an isomorphism’ between two structural equations models? A function, f , from the values of the variables in \mathcal{M}_1 to the values of the variables in \mathcal{M}_2 is an *isomorphism* iff (a) it is a bijection, (b) it preserves the mappings of the structural equations, and (c) it maps actual values to actual values. That is: (a) different variable values in \mathcal{M}_1 get mapped to different variable values in \mathcal{M}_2 , and every variable value in \mathcal{M}_2 has some variable value in \mathcal{M}_1 which is mapped to it; (b) for every structural equation in \mathcal{M}_1 , if that equation maps u_1, u_2, \dots, u_N to v , then there is an equation in \mathcal{M}_2 which maps $f(u_1), f(u_2), \dots, f(u_N)$ to $f(v)$; and (c) if v is an actual variable value in \mathcal{M}_1 , then $f(v)$ is an actual variable value in \mathcal{M}_2 .
6. You may instead want to contend that there’s something wrong with one of the structural equations models we’ve written down here. See Blanchard & Schaffer (2017) for this kind of response. But notice that, unlike the kinds of cases which Blanchard & Schaffer discuss, the values of the variables in *Preemptive Overdetermination* and *Tornado* do not seem to correspond to possibilities which we are ‘not willing to take seriously’. These structural equations models do impose an artificial and unrealistic simplicity on scenarios which, in any realistic case, would be much messier, involve more variables, and perhaps be indeterministic. But I don’t see why this should matter. It is not cognitively taxing to make the unrealistic assumption that matters really are this simple and deterministic. (If anything, it is making *realistic* assumptions which would be cognitively demanding.) So it’s unclear why we should doubt the intuitions we have once we’ve made these simplifying assumptions. Relatedly, we may imagine simplistic systems of neurons which have structures similar to *Preemptive Overdetermination* and *Tornado*. These simplistic causal systems can be correctly modelled by isomorphic systems of equations without the need of any simplifying assumptions. See Gallow (2021, §1.1) and Blanchard & Schaffer (2017, fn 23).

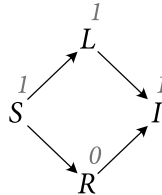
For another illustration of the same issue, consider the following vignette:

Switch

The power cord for the light bifurcates into a left wire and a right wire, and there is a switch with only two positions: Left and Right. If the switch is set to Left then it will direct any current along the left wire. If the switch is set to Right, then it will direct any current along the right wire. Both the left and right wires attach to the light, and if current is flowing through either wire, the light will be illuminated. In the morning, Filipa flips the switch from Right to Left. In the evening, Phoebe turns on the power. Current then flows through the left wire, and the light turns on.⁷

Let's focus on the following variables: whether the switch is set to Left ($S = 1$) or Right ($S = 0$), whether current is flowing through the left wire ($L = 1$) or not ($L = 0$), whether current is flowing through the right wire ($R = 1$) or not ($R = 0$), and whether the light is illuminated ($I = 1$) or not ($I = 0$). And let us suppose that these variables influence each other in the way described by this system of structural equations:

$$\begin{aligned} I &:= L \vee R \\ L &:= S \\ R &:= \neg S \\ S &= 1 \end{aligned}$$



It does not appear that the light being illuminated ($I = 1$) is an effect of the switch being set to Left ($S = 1$), rather than Right ($S = 0$). The switch makes a difference to whether current flows through the left or the right wire, but it does not make any difference to whether the light is illuminated or not. Relatedly, there seems to be an important difference between Filipa and Phoebe. While Phoebe can take credit for the light being illuminated, Filipa cannot. Filipa can at most take credit for current flowing through the left wire, rather than the right.

However, once again, this system of structural equations is isomorphic to the one we wrote down for *Preemptive Overdetermination*. We may associate the variable S with M , L with P , R with E , and I with D . Then, ‘corresponding’

⁷ See Hall (2000), Pearl (2000, example 10.3.6), Halpern & Pearl (2005), and Sartorio (2005, 2016), a.o.

variables are related by exactly the same equations and take on exactly the same values. So any theory of causation formulated in terms of structural equations models alone will not distinguish between *Preemptive Overdetermination* and *Switch*.

Several authors have responded to observations like these by introducing a distinction between variable values which represent default or inertial states and those which represent departures from a default or inertial state.⁸ Just to have a name, I'll call a deviation from the default *deviant*.

I won't have the space to say very much about this difference between variable values which are default and those which are deviant. But just to help the reader acquire a familiarity with the distinction, let me offer the following rough-and-ready characterisation. If it feels natural to describe a variable value by saying 'nothing happened'—or if it is natural to describe it as representing a *state*, as opposed to an *event*—then that variable value is likely default. On the other hand, if it feels natural to describe a variable value by saying that something *happened*, or that it represents an *event*, then that variable value is likely a deviant departure from the default. We tend to expect that the states of the world represented by default variable values will persist so long as they are left alone. Another helpful characterisation can be given in terms of what we are inclined to imagine when we counterfactually suppose that some event did not take place. When we counterfactually suppose that the chef didn't poison the president's food, we're not inclined to imagine the chef poisoning his drink, or shooting the president with a revolver, or staging a production of *West Side Story*. Instead, we imagine him preparing food without poison. If a variable value represents a state which we're inclined to imagine when counterfactually supposing an event away, then it is likely a more default value. For this reason, we tend to be unsure how to counterfactually imagine away default variable values when there are multiple possible contrasts. When asked to counterfactually suppose the chef didn't poison the president's food, we easily imagine him preparing unpoisoned food. In contrast, consider Ali, who is just standing about, doing nothing. If you're asked to suppose that Ali isn't just standing about, doing nothing, it's unclear what you're being asked to imagine. No scenario springs to mind as being the kind of thing you're meant to be coun-

8. See Hart & Honoré (1985), Kahneman & Miller (1986), Thomson (2003), Maudlin (2004), Menzies (2004, 2007, 2017), McGrath (2005), Hall (2007), Hitchcock (2007), Halpern (2008, 2016), Hitchcock & Knobe (2009), Paul & Hall (2013), Halpern & Hitchcock (2015), Wolff (2016), Icard *et al.* (2017), Gerstenberg & Icard (2020), and Gallow (2021), a.o. For criticism of this response, see Blanchard & Schaffer (2017) and Wysocki (ms).

terfactually supposing to be the case, indicating that Ali standing about doing nothing is a default.⁹

When variables have multiple values, we shouldn't work with a binary distinction between variable values which are default and those which are deviant. Instead, we should order the values of variables in terms of which are more default than which others. Turn again to the chef from *Preemptive Overdetermination*, and consider a variable, C^* , which takes on three potential values: $C^* = 0$ if the chef just stands about, doing nothing. $C^* = 1$ if the chef prepares a normal meal. And $C^* = 2$ if the chef prepares a poisoned meal. If I ask you to counterfactually suppose that the chef didn't prepare a poisoned meal, it's most natural to imagine him preparing a normal, unpoisoned meal. This suggests that $C^* = 1$ is more default than $C^* = 2$. And, if I ask you to counterfactually suppose that the chef didn't prepare a normal meal, it's most natural to imagine him standing about, doing nothing (especially in a context where we haven't raised the possibility of the chef poisoning the meal). This suggests that $C^* = 0$ is more default than $C^* = 1$.

Distinguishing (more) default variable values from (more) deviant ones allows us to distinguish *Preemptive Overdetermination* from *Tornado*. For, while the variable value $C = 0$ represents the default state of the CIA agent doing nothing, the 'corresponding' variable value $\overline{D} = 0$ represents the house's destruction, which is a deviation from the default state of the home remaining intact. Likewise, this distinction allows us to distinguish *Preemptive Overdetermination* from *Switch*. For $M = 0$ is more default than $M = 1$ —that is, the MI6 agent doing nothing is more default than their providing poison to the chef. However, $S = 0$ is no more default than $S = 1$. It's no more default for the switch to be set to Left than it is for the switch to be set to Right. (It's natural to describe both the switch being set to Left and the switch being set to Right as *states*, rather than *events*, and both of these states are inertial ones in which the switch will remain so long as it is left alone. If I ask you to counterfactually suppose that the switch is not set to Left, you'll naturally imagine that it's set to Right; but likewise, if you're asked to suppose it's not set to Right, you'll equally naturally imagine that it's set to Left. So neither value appears any more default than the other.)

9. Compare with Hall (2007, §6) and Hitchcock (2007, §5).

4 Rules for Causal Process Tracing

In *Preemptive Overdetermination*, we can trace a process from the MI6 agent's providing poison ($M = 1$), to the chef poisoning the food ($P = 1$), to the president's death ($D = 1$). We can trace the process from $M = 1$ to $P = 1$ because, when we wiggle M 's value from 1 to 0, P 's equation tells us that P changes from 1 to 0. And we can trace the process from $P = 1$ to $D = 1$ because, when we wiggle P 's value from 1 to 0, D 's equation tells us that D changes from 1 to 0.

In general, tracing a process from C to one of its children, D , involves changing C 's value in D 's equation, and seeing how D 's value changes in response. When a variable is binary, it's clear how to change its value. There is only one alternative value for the variable to take on. But when variables take on more than two values, we should be more careful and explicitly stipulate which contrast value we are changing C to. Suppose that, when we change C 's value from c to $c^* \neq c$ in D 's equation, D 's value changes from d to $d^* \neq d$. In that case, I'll say that D taking on the value d , rather than d^* , is a *consequence* of C taking on the value c , rather than c^* .¹⁰ As a notational convention, I'll write ' $(v, v^*)_V$ ' for the variable V 's taking on the value v , rather than $v^* \neq v$. So, if the value of D determined by D 's equation changes from d to d^* when we change C 's value from c to c^* , I'll say that $(d, d^*)_D$ is a consequence of $(c, c^*)_C$. And, in general, I'll refer to any variable value, contrast pair $(v, v^*)_V$ as a consequence for V .¹¹

Causal process tracing is a matter of tracing out consequences like this. You begin with some collection of variables values, $C_1 = c_1, C_2 = c_2, \dots, C_N = c_N$, along with a range of contrast values, $c_1^* \neq c_1, c_2^* \neq c_2, \dots, c_N^* \neq c_N$. This gives you a collection of 'initial' consequences $(c_1, c_1^*)_{C_1}, (c_2, c_2^*)_{C_2}, \dots, (c_N, c_N^*)_{C_N}$. From there, you can start to trace out further consequences, according to the following rules.

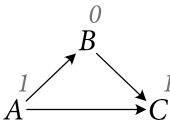
Rule #0: How to Trace Consequences If U_1, U_2, \dots, U_N are all parents of V , then it's possible to trace a consequence for V , $(v, v^*)_V$, from the consequences $(u_1, u_1^*)_{U_1}, (u_2, u_2^*)_{U_2}, \dots, (u_N, u_N^*)_{U_N}$, iff, when you change

10. For more of the rule of contrasts in causal claims, see Hitchcock (1996a,b, 2011), Maslen (2004), Schaffer (2005, 2012a), and Gallow (2021, §5.1), a.o.

11. In particular, I'll refer to $(c, c^*)_C$ as a consequence for C , even when $(c, c^*)_C$ initiates the causal process, and so is not a consequence of anything else in the process. This terminological stipulation is slightly awkward, but it allows me to concisely state the recursive rules #0, #2, #3, and #4.

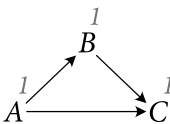
the value of every U_i from u_i to u_i^* in V 's equation—and leave the value of every other variable unchanged— V 's value changes from v to $v^* \neq v$.

The zeroth rule is more of a definition. It explains what it means to extend a causal process by tracing one consequence from some others which you have already included in the process. It's natural to think of rule #0 as saying that you can extend a process from $(u_1, u_1^*)_{U_1}, (u_2, u_2^*)_{U_2}, \dots, (u_N, u_N^*)_{U_N}$ to $(v, v^*)_V$ whenever the latter consequence counterfactually depends upon the former consequences. However, given the usual structural equations model semantics for counterfactuals,¹² this isn't quite correct. To appreciate why, consider the system of structural equations shown below.

$$\begin{aligned} C &:= A \vee B \\ B &:= \neg A \\ A &= 1 \end{aligned}$$


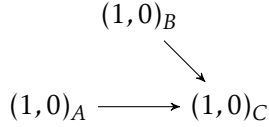
In this system of equations, rule #0 tells us that we may trace a process from $(1, 0)_A$ to $(1, 0)_C$. For, when we look at the variable values appearing in C 's equation, we have $C := 1_A \vee 0_B = 1$ (I've subscripted the variable values to make it clear which variables they are values of). And, if we change 1_A to 0_A without changing the value of B , we get that $C := 0_A \vee 0_B = 0$. So, just looking at C 's equation, changing A 's value from 1 to 0 changes C 's value from 1 to 0. But $C = 1$ does not counterfactually depend upon $A = 1$. For, if A were to be 0, B would have been 1, and so C would have remained 1. Just to have a term to mark this distinction, we could say that, while $(1, 0)_C$ does not *globally* depend upon $(1, 0)_A$, $(1, 0)_C$ does *locally* depend upon $(1, 0)_A$. In these terms, rule #0 tells us that we may extend a process by tracing $(v, v^*)_V$ from $(u_1, u_1^*)_{U_1}, (u_2, u_2^*)_{U_2}, \dots, (u_N, u_N^*)_{U_N}$ whenever the former consequence locally depends upon the latter consequences.

The rules for causal process tracing do not require us to begin tracing from a single initial consequence, $(c, c^*)_C$. We may instead, if we wish, start tracing from a collection of consequences. Consider, for instance, the following structural equations model:

$$\begin{aligned} C &:= A \vee B \\ B &:= A \\ A &= 1 \end{aligned}$$


12. See Galles & Pearl (1998), Hiddleston (2005a), and Briggs (2012).

In this model, we may begin with the consequences $(1, 0)_A$ and $(1, 0)_B$, and extend the process by tracing out the consequence $(1, 0)_C$.



Rule #1 is the first substantive rule. It says something about the order in which you must trace out consequences when you are extending a causal process. Roughly, the rule says that, before you decide whether to include a consequence for a variable, V , in the process, you must first have decided whether to include a consequence for any parent of V in the process. You cannot first include a consequence for V and only later include a consequence for one of V 's parents. More carefully, let's say that one variable, U , is *closer* to an initial variable, C , than V is iff (a) there is a directed path from C to U and a directed path from C to V , and (b) the longest directed path from C to U is shorter than the longest directed path from C to V .

Rule #1: Trace Out Consequences in Order If U is closer to some initial variable than V is, then you must decide whether to include a consequence for U in the process before you decide whether to include a consequence for V .

This rule prevents us from first including a consequence for V , and only later including a consequence for one of V 's parents. For, if U is a parent of V which lies downstream of an initial variable, then U must be closer to that initial variable than V is.

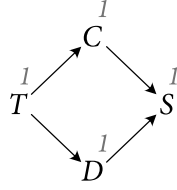
Rule #1 is important because of rule #2, which says that, if you're going to trace a process forward to a consequence for a variable, V , then it must be a consequence of all of the consequences for V 's parents which you've already included in the process.

Rule #2: Trace Forward From All Consequences You may only include $(v, v^*)_V$ in a causal process if you can trace it forward from *all* of the consequences for V 's parents that you've included in the process.

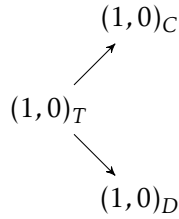
For instance, consider the structural equations model we wrote down for *Tornado*, but let's exchange the variable \overline{D} for the less confusing variable D , which

is 1 if the house is destroyed and 0 if it is not.

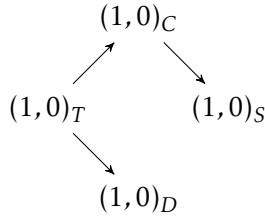
$$\begin{aligned} S &:= C \vee \neg D \\ C &:= T \\ D &:= T \\ T &= 1 \end{aligned}$$



Suppose we've already traced out a process leading from the tornado, $(1, 0)_T$, to Aunt Em's being in the storm cellar, $(1, 0)_C$, and to the house's being destroyed, $(1, 0)_D$.



At this point, we may *not* trace this process forward from Aunt Em's being in the storm cellar, $(1, 0)_C$ to her survival, $(1, 0)_S$. This is *not* a causal process:



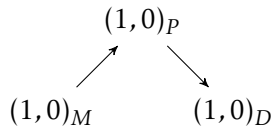
It is true that Aunt Em's survival depends upon her being in the storm cellar. That is: looking at S 's equation, changing C from 1 to 0 changes S 's value from 1 to 0. But rule #2 tells us that any consequence for S which we're going to include in this process must be a consequence of both $(1, 0)_C$ and $(1, 0)_D$. But, looking at the variable values in S 's equation, $S := 1_C \vee \neg 1_D = 1$, changing both C and D from 1 to 0 gives us $S := 0_C \vee \neg 0_D = 1$. So rule #2 tells us that we cannot include any consequence for S in this process.

The third rule is the one which allows us to distinguish the case of *Preemptive Overdetermination* from *Tornado*. Let's say that the consequence $(v, v^*)_V$ is *deviant* iff v^* is more default than v ; otherwise, we can say that $(v, v^*)_V$ is *default*. Then, rule #3 tells us that, while you are sometimes permitted to exclude default consequences from a causal process, you are never permitted to exclude deviant consequences. If the other rules allow you to include a de-

viant consequence for the variable V in the causal process you are tracing, this consequence *must* be included.

Rule #3: Trace All Deviant Consequences If a consequence $(v, v^*)_V$ is deviant and the other rules allow you to include it in the process, then it must be included.

In *Preemptive Overdetermination*, we may trace out a process from the MI6 agent’s providing poison, rather than not, $(1, 0)_M$, to the chef’s poisoning the food, rather than not, $(1, 0)_P$, to the president’s dying, rather than remaining alive, $(1, 0)_D$.



We *could* have traced the consequence $(0, 1)_E$ from $(1, 0)_M$, since the CIA agent’s not igniting their explosives depends upon the MI6 agent providing poison to the chef. If we had included this consequence in the process, we would not have been able to trace out any consequences for whether the president died. However, the consequence $(0, 1)_E$ is default. That is: the CIA agent doing nothing is more default than their igniting explosives in the presidential palace. And if $(0, 1)_E$ is default, then we have the option of not including it in the causal process we are tracing. So we have the option of tracing a causal process from the MI6 agent’s action to the president’s death.

In contrast, in *Tornado*, the destruction of the house is not default. It is an event, a happening. And the house remaining intact is an inertial, default state. So $(1, 0)_D$ is a deviant consequence. Rule #3 therefore requires us to include $(1, 0)_D$ in any causal process we trace from $(1, 0)_T$. If we include this consequence, rule #2 tells us that any consequence for S must depend upon $(1, 0)_D$. But in S ’s structural equation, $S := C \vee \neg D$, $D = 0$ is sufficient for $S = 1$. So changing the value of D from 1 to 0 will not change the value of S —whether we include the consequence $(1, 0)_C$ or not. So Aunt Em’s survival is not an effect of the tornado.

You may always trace out a causal process by including all the deviant consequences and excluding any default consequences. While deviant consequences are mandatory, you always have the option of excluding every default consequence. However, once you include *some* default consequence for a variable, U , $(u, u^*)_U$, from that point on, you must continue to trace out every possible consequence downstream of $(u, u^*)_U$ —at least, you must continue tracing out

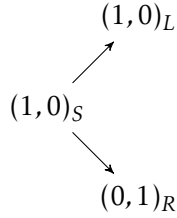
every consequence of $(u, u^*)_U$ until those consequences have resolved themselves back into a single deviant consequence. At that point, you may once again begin tracing only deviant consequences, if you wish. Think about it like this: there are two kinds of causal processes, which we can call *productive* processes and *dependence* processes.¹³ To trace out a productive process, you must include all and only the deviant consequences. But if you want to trace out a dependence process, then you must include *every* consequence, even the default ones. (We'll see in section 5 below that including every consequence is equivalent to checking for counterfactual dependence, whence the name 'dependence process'.) If you start out tracing a productive process, you may, at any point you wish, switch over to tracing out a dependence process by including some default consequence. However, from that point forward, you must continue tracing out the dependence process until it has resolved itself into a single consequence.

To state that a bit more carefully, let's say that a default consequence, $(u, u^*)_U$, has been *resolved* into a single consequence just in case there's some descendant of U 's, R (for 'resolved'), such that, for every variable, V , which is a descendant of U but not a descendant of R , either (a) you have included a consequence for V in the process or (b) the rules do not allow you to include a consequence for V in the process. Then, the fourth rule says that, once you include a default consequence $(u, u^*)_U$ in a process, you must continue tracing out every consequence downstream of $(u, u^*)_U$ until it is resolved.

Rule #4: Trace All Consequences of Unresolved Default Consequences If V is a descendant of an unresolved default consequence and the other rules allow you to include $(v, v^*)_V$ in the process, then it must be included.

Rule #4 is relevant when we consider cases like *Switch*. If we are interested in the effects of the switch being set to Left, rather than Right, then we will have to begin tracing a causal process with the default consequence $(1, 0)_S$. Once we've included this default consequence in our causal process, we must continue tracing all possible consequences downstream of it until it is resolved. This means that we must include both the deviant consequence that the left wire has current running through it, $(1, 0)_L$, and the default consequence that the right wire has no current flowing through it, $(0, 1)_R$.

13. Compare with Hall (2004).



But once we've included both of these consequences in our causal process, we cannot trace out any consequences for whether the light is illuminated. When we look at the variable values in I 's equation, $I := 1_L \vee 0_R = 1$, changing L from 1 to 0 and changing R from 0 to 1 gives us $I := 0_L \vee 1_R = 1$. So the light's being illuminated is not an effect of the switch being set to Left, rather than Right.

These are all the rules for causal process tracing. Any process you are able to trace out in accordance with these rules is a causal process. Causation is then defined in terms of causal processes. But there are two subtle questions to address. Firstly: what are the relata of the causal relation? We might decide to say that they are what I have here called *consequences*. That is, we may want to say that E taking on the value e , rather than e^* , is an effect of the initial variables, C , taking on their actual values, \mathbf{c} , rather than some contrast values, \mathbf{c}^* .¹⁴ On this view, causation is a binary relation between pairs of variable values (or collections of variable values).¹⁵ Alternatively, we might want to say that causation is a binary relation between variable values (or collections thereof), and that $C = \mathbf{c}$ is a cause of $E = e$ if, for *some* contrasts \mathbf{c}^* and e^* , we can trace an appropriate causal process from $(\mathbf{c}, \mathbf{c}^*)_C$ to $(e, e^*)_E$. From my perspective, there's little to tell between these two approaches. Sometimes our causal claims make explicit reference to contrasts, which may be thought to favour the first approach. However, most often our causal claims do not involve any explicitly stated contrasts; and it's unclear whether the function of mentioning contrasts is to specify the relata of the causal relation or to instead specify the causal process linking cause to effect. I'm going to opt for the second view here, but a reader who prefers the first can accept everything else I'll have to say with a few minor and superficial changes.

14. Here, I'm using the boldface ' C ' for a set of variables, and I'm using ' \mathbf{c} ' and ' \mathbf{c}^* ' for assignments of values to those variables. That is, \mathbf{c} and \mathbf{c}^* are functions from the variables in C to their values. \mathbf{c} maps every $C \in C$ to its actual value, and \mathbf{c}^* maps every $C \in C$ to some non-actual value.
15. Some might prefer to say that causation is a four-place relation between variable values, with the first and third places occupied by C and E 's actual values, and the second and fourth places occupied by contrast values for C and E . See Maslen (2004) and Schaffer (2005). I believe this is just a notational variant of the view I propose in the body.

Secondly: when can we use a causal process to conclude that $E = e$ is an effect of $C = c$? Whenever (i) the process is initiated by consequences for the variables in C ; (ii) a consequence for E is included in the process; and (iii) the causal process is *minimal*.

Causation If you can trace a minimal causal process from $(c, c^*)_C$ to $(e, e^*)_E$, then $C = c$ is a cause of $E = e$.

If $C = c$ is a cause of $E = e$, then I'll say that, for any $C \in C$, C 's value is a *part* of a cause of E 's value. It, together with the other variables in C , brought about $E = e$. A causal process from $(c, c^*)_C$ to $(e, e^*)_E$ is *minimal* just in case there is no sub-process of it leading from $(\tilde{c}, \tilde{c}^*)_{\tilde{C}}$ to $(e, e^*)_E$ —where $\tilde{C} \subseteq C$ and \tilde{c}^* gives the same contrasts to the variables in \tilde{C} as c^* does—which is also a causal process.

To see why the minimality condition is important, let's extend our model of *Switch* by including a variable, P , for whether the power is on. If the power is on, then $P = 1$; whereas, if the power is off, $P = 0$. Then, we have the following system of structural equations.

$$\begin{array}{l}
 I := L \vee R \\
 L := S \wedge P \\
 R := \neg S \wedge P \\
 S = P = 1
 \end{array}
 \qquad
 \begin{array}{c}
 \begin{array}{ccc}
 S & \xrightarrow{1} & L \\
 & \searrow & \nearrow \\
 & & I \\
 P & \xrightarrow{1} & R \\
 & \nearrow & \searrow \\
 & & I
 \end{array}
 \end{array}$$

The rules allow us to trace out the following causal process:

$$\begin{array}{c}
 (1, 0)_S \\
 \searrow \\
 (1, 0)_L \longrightarrow (1, 0)_I \\
 \nearrow \\
 (1, 0)_P
 \end{array}$$

But we should not conclude that the switch being set to Left was a part of a cause of the light being illuminated—we shouldn't say that the switch and the power together caused the light to be illuminated. For the consequence $(1, 0)_S$ is an inessential part of this causal process. Even without it, we can trace the causal process

$$(1, 0)_P \longrightarrow (1, 0)_L \longrightarrow (1, 0)_I$$

This second causal process is a sub-process of the first one. So the first causal process is not minimal.

5 Further Discussion

In this section, I will apply the theory of causation from section 4 to some additional examples and note some of its properties.

In the first place, note that, if $E = e$, rather than e^* , (globally) counterfactually depends upon $C = c$, rather than c^* , then there will be a causal process leading from $(c, c^*)_C$ to $(e, e^*)_E$. In fact, there will be a *dependence* process leading from $(c, c^*)_C$ to $(e, e^*)_E$ —a process which includes every possible consequence. In general, $E = e$, rather than e^* , counterfactually depends upon $C = c$, rather than c^* , if and only if there is a dependence process from $(c, c^*)_C$ to $(e, e^*)_E$.¹⁶ So, according to this theory, counterfactual dependence is sufficient for causation, in the following sense: if $E = e$ counterfactually depends upon $C = c$, then the values of some subset of C is a cause of $E = e$.

This means that, in particular, the theory recognises cases of prevention and ‘double prevention’ as causation.¹⁷ Consider, for instance:

Jewel Heist

James breaks into the museum to steal the jewels. An alarm would have prevented him from getting the jewels, but last night, Dalton disabled the security system, preventing the alarm from preventing James from stealing the jewels.

Using ‘ J ’ for whether James steals the jewels, ‘ A ’ for whether the alarm goes off, and ‘ D ’ for whether Dalton disables the security system, let’s suppose that these variables influence each other in the way described by the structural equations

16. In a dependence process emanating from $(c, c^*)_C$, every variable’s contrast is the value it would take on, were each $C \in C$ to take on the value from c^* . We can show this by induction on a variable’s ‘distance’ from C , where V ’s distance from C is the length of the longest directed path from some $C \in C$ to V . The base case, where the distance is 0, is trivial. (The $C \in C$ are the only variables a distance of 0 from C .) So suppose that it holds for every variable whose distance from C is at least k . Then, take any variable, V , whose distance from C is $k + 1$. Each of V ’s parents is either not a descendant of any $C \in C$, in which case it would take on its actual value, were C to be c^* , or else V ’s parent is closer to C than V , in which case its contrast in the dependence process is the value it would take on, were C to be c^* (by the inductive hypothesis). By rules #0, #1, and #2, V ’s contrast in the process must be the value determined by its structural equation, when the parents ‘on the process’ are given their contrasts and the parents ‘off the process’ are given their actual values. Since these are also the values the parents would take on, were C to be c^* , V ’s contrast will be the value V would take on, were C to be c^* . V was arbitrary, so the same goes for every other variable a distance of $k + 1$ from C .

17. Cases of ‘double prevention’ are discussed in Hall (2004). See also Schaffer (2000, 2012b).

below.

$$\begin{array}{l}
 J := \neg A \\
 A := \neg D \\
 D = 1
 \end{array}
 \qquad
 \begin{array}{c}
 \overset{1}{D} \longrightarrow \overset{0}{A} \longrightarrow \overset{1}{J}
 \end{array}$$

Then, we can trace out the causal process $(1, 0)_D \rightarrow (0, 1)_A \rightarrow (1, 0)_J$. So Dalton’s disabling the security system prevented the alarm from sounding. (That is to say: the alarm not going off is an effect of Dalton’s disabling the security system.) And, the alarm would have prevented the jewel heist. Since Dalton prevented this potential preventer (‘double prevention’), the jewel heist is another effect of him disabling the security system.

The theory also recognises cases of prevention and ‘double prevention’ without (global) counterfactual dependence. Consider, for instance,

Preemptive Prevention

As in *Jewel Heist*, except that, if Dalton hadn’t disabled the security system, then Brynn (the ‘backup’) would have cut the power to the museum, and the alarm still wouldn’t have gone off.

To model this version of the case, we can use ‘*B*’ for whether Brynn cuts the power, and continue to use ‘*D*’, ‘*A*’, and ‘*J*’ in the same way. And we can assume that the relations of influence between these variables are as described by the equations below.

$$\begin{array}{l}
 J := \neg A \\
 A := \neg D \wedge \neg B \\
 B := \neg D \\
 D = 1
 \end{array}
 \qquad
 \begin{array}{c}
 \overset{1}{D} \longrightarrow \overset{0}{A} \longrightarrow \overset{1}{J} \\
 \searrow \quad \nearrow \\
 \quad \overset{0}{B}
 \end{array}$$

Brynn’s failure to cut the power is a default consequence, so it need not be traced, and we can trace out the same causal process as in *Jewel Heist*, $(1, 0)_D \rightarrow (0, 1)_A \rightarrow (1, 0)_J$. So the theory tells us that both the failure of the alarm to sound and the success of the jewel heist are effects of Dalton’s disabling the security system. Dalton deserves credit for the alarm not sounding, even though the ‘back up’ Brynn means that the alarm’s silence doesn’t (globally) counterfactually depend upon what Dalton did.¹⁸

18. Cases called ‘preemptive prevention’ are discussed in McDermott (1995) and Collins (2004). However, the cases they discuss are somewhat different from the case I am calling *Preemptive Prevention*. In the McDermott and Collins cases, the prevented event does not even locally depend upon the preempting preventer; whereas, in *Preemptive Prevention*, the failure of the

To my knowledge, no other extant theory of causation generates this collection of verdicts. Some deny that there is causation in cases like *Preemptive Overdetermination*.¹⁹ Some say that the tornado caused Aunt Em to survive, and that the switch's being set to Left caused the light to be illuminated.²⁰ Some deny that there can be causation by prevention or 'double prevention'.²¹ Other theories fail to issue any verdicts at all about some of our cases.²²

Preemptive Prevention is a case in which the theory from section 4 disagrees with the theory of causation I provided in Gallow (2021). For another case in which the theories disagree, consider the following vignette:

Coordination Game

Fozzie Bear and Crazy Harry play a coordination game. Each has a switch with two positions: Left and Right. Fozzie has first move. He can either flip his switch or leave it alone. Next, after learning what Fozzie has done, Crazy Harry can either flip his switch or leave it alone. If their switches are aligned, they win \$1,000,000. If their switches are misaligned, the money will be incinerated in an extravagant explosion. To start, the switches are misaligned: Fozzie's is set Left and Harry's is set Right. Wanting the money, Fozzie flips his switch to Right. Seeing this and wanting the explosion, Harry flips his switch to Left. The money is incinerated.²³

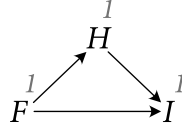
Let's use 'F' for whether Fozzie Bear flips his switch, 'H' for whether Crazy Harry flips his switch, and 'I' for whether the money is incinerated. Then, the relations of influence between these variables are described by these structural

alarm to sound does locally depend upon Dalton disabling the security system. The McDermott and Collins cases are more similar to a version of *Preemptive Prevention* where Dalton disables the security system and Brynn cuts the power anyway. (In that version of the case, Brynn would be the 'preempting' preventer.)

19. For instance, Mackie (1965), Suppes (1970), Eells (1991), Beckers & Vennekens (2017, 2018), and Andreas & Günther (forthcoming).
20. For instance, Lewis (1973, 1986, 2004), Ramachandran (1997), Schaffer (2001), Hitchcock (2001), Woodward (2003), Yablo (2004), Halpern & Pearl (2001, 2005), Hall (2007) (see Hitchcock, 2009), Halpern (2016), Andreas & Günther (2020, 2021), and Bochman (2021).
21. For instance, Aronson (1971), Fair (1979), Salmon (1984, 1994), Ehring (1997), and Dowe (2000).
22. For instance, Hitchcock (2007) and Weslake (forthcoming).
23. This case is modelled on McDermott (1995)'s *Shock C*.

equations.

$$\begin{aligned} I &:= [F = H] \\ H &:= F \\ F &= 1 \end{aligned}$$



(Here, $[F = H]$ is the truth-value of $F = H$. It is 1 if F and H have the same value, and 0 otherwise.)

If we assume that flipping your switch is less default than doing nothing, and that money being incinerated is less default than the money remaining intact, then the theory I provided in Gallow (2021) allows us to trace a causal process from Fozzie Bear’s flipping his switch, $(1, 0)_F$, to Crazy Harry’s flipping his, $(1, 0)_H$, to the money being incinerated, $(1, 0)_I$. This seems like a bad result, since it seems wrong to say that the money being incinerated was an effect of Fozzie’s flipping his switch. We are instead inclined to say that, given that Harry wanted the explosion, and was going to flip his switch iff Fozzie flipped his, Fozzie’s flip didn’t make any difference to whether the money was incinerated. On the present theory, $(1, 0)_F \rightarrow (1, 0)_H \rightarrow (1, 0)_I$ is not a causal process, since it violates rule #2. The consequence $(1, 0)_I$ is only traced forward from $(1, 0)_H$, even though both F and H are parents of I . Any consequences for whether the money is incinerated would have to be traced forward from both Fozzie’s flipping his switch and Harry’s flipping his. But the money’s incineration does not depend upon both Fozzie and Harry’s flips. Had neither of them flipped their switches, the money would still have been incinerated. Nor is $(1, 0)_F \rightarrow (1, 0)_I$ a causal process, for it does not include the deviant consequence $(1, 0)_H$, in violation of rule #3.

In Gallow (2021), I was trying to provide a theory of causation which satisfied the following principle:

Invariance under Interpolated Variable Removal If $V \neq C, E$ is interpolated along a path in the structural equations model \mathcal{M} , then a theory of causation should tell you that $C = c$ is a cause of $E = e$ in \mathcal{M} iff it tells you that $C = c$ is a cause of $E = e$ in \mathcal{M}^{-V} .

This requires some explanation. I say that the variable V is *interpolated along a path* in a structural equations model iff it has a single parent, Pa , a single child, Ch ,

$$Pa \rightarrow V \rightarrow Ch$$

and Pa is not also a parent of Ch . If V is interpolated along a path in \mathcal{M} ,

then \mathcal{M} contains an equation of the form $V := f(Pa)$, for some function f of Pa . And it contains a structural equation of the form $Ch := g(\dots V \dots)$, for some function, g , of V and perhaps some other variables. \mathcal{M}^{-V} is the model \mathcal{M} with the variable V removed. If V is interpolated along a path in \mathcal{M} , then, to get the model \mathcal{M}^{-V} , you just get rid of the equation $V := f(Pa)$ entirely, and replace the equation $Ch := g(\dots V \dots)$ with $Ch := g(\dots f(Pa) \dots)$. The principle tells us that whether you interpolate a variable along a path or not, this shouldn't make any difference to what your theory tells you about the causal relations between the other variables in the model.

I've come to think that this principle is too strong, in part because it requires us to treat *Coordination Game* like a case of preemptive overdetermination. Consider the following vignette:

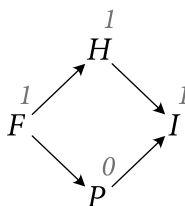
Preemptive Overdetermination (v2)

Kermit the Frog and Miss Piggy both have switches in front of them, with two positions: Off and On. To start, both switches are Off. Kermit has first move: he can either flip his switch to On or leave it alone. Next, after learning what Kermit has done, Miss Piggy can either flip her switch to On or leave it alone. If either switch is On, then a corresponding bomb will be activated. If exactly one of the bombs is activated, then \$1,000,000 will be incinerated by the exploding bomb. If both bombs are activated, there will be a power surge and neither bomb will go off. Both players know this, and both players just want to watch the money burn. So Kermit flips his switch to On, and Miss Piggy does nothing. (Had Kermit not flipped, Miss Piggy would have.) Kermit's bomb is activated and Miss Piggy's is not. Kermit's bomb explodes, incinerating the money.

In my opinion, this vignette is in all relevant respects just like *Preemptive Overdetermination*. The CIA agent and Miss Piggy are backup, would-be causes of the president's death and the money's incineration, respectively. These backup, would-be causes are preempted by the MI6 agent and Kermit the Frog, respectively. Just as the MI6 agent caused the president to die, Kermit caused the money to be incinerated.

We can model *Preemptive Overdetermination (v2)* with the following binary variables: I , for whether the money is incinerated, P , for whether Miss Piggy's bomb is activated, H , for whether Kermit's bomb is activated, and F , for whether Kermit the Frog flips his switch. These variables influence each

other in the ways described by the equations below.

$$\begin{aligned}
 I &:= [\neg P = H] \\
 P &:= \neg F \\
 H &:= F \\
 F &= 1
 \end{aligned}$$


Notice that, in this structural equations model, the variable P is interpolated along a path. According to *Invariance under Interpolated Variable Removal*, we can remove it, and this won't make any difference to what our theory has to tell us about whether $F = 1$ is a cause of $I = 1$. But removing the variable P leaves us with this model:

$$\begin{aligned}
 I &:= [F = H] \\
 H &:= F \\
 F &= 1
 \end{aligned}$$


And this is exactly the model we wrote down for *Coordination Game*. Moreover, there doesn't seem to be any difference with respect to which values of the corresponding variables are default and which are deviant. In both cases, $I = 1$ stands for the deviant event of \$1,000,000 being incinerated, while $I = 0$ stands for the default, inertial state of the money remaining intact. In *Coordination Game*, $H = 1$ stands for the deviant event of Harry flipping his switch, and $H = 0$ stands for the default, inertial state of Harry doing nothing. Whereas, in *Preemptive Overdetermination (v2)*, $H = 1$ stands for the deviant event of Kermit's bomb being activated, and $H = 0$ stands for the default, inertial state of the bomb remaining deactivated. And, in both cases, $F = 1$ stands for the deviant event of someone (either Fozzie Bear or Kermit the Frog) flipping a switch, and $F = 0$ stands for the default, inertial state of them doing nothing. So it doesn't seem that we can use differences in which variable values are default and which are deviant to distinguish these two structural equations models.

One reaction to this observation is to think that we were wrong to think that Fozzie Bear didn't cause the money to be incinerated. For instance, we might suspect that our intuitions are being misled by the fact that, while Kermit *intended* to incinerate the money, Fozzie Bear did not. That's an important asymmetry between Kermit and Fozzie, and it's easy to understand how this might influence our causal judgements. For you may think that, in general, we have a tendency to conflate causal and moral responsibility, and that intentions

are relevant to moral responsibility. This error theory is *prima facie* plausible, but unfortunately, I don't think that it stands up to scrutiny. In the first place, notice that changing Kermit's intentions doesn't seem to affect the intuition that he caused the money to be incinerated. Suppose Kermit was trying to save the money, but was under the false impression that flipping the switch would deactivate his bomb. In this version of the case, it appears that Kermit *unwittingly* caused the money to be incinerated, not that he didn't cause the incineration. In the second place, it still seems like Fozzie Bear didn't cause the money to be incinerated when we change *his* intentions. Suppose, for instance, that Fozzie wants the explosion, but he's misinformed, and think that this will only happen if the switches are aligned. (Harry is not misinformed.) So Fozzie flips his switch with the intention of incinerating the money, Harry undoes this blunder by flipping his switch, and the money is incinerated. In this version of the case, it still seems like Fozzie Bear's flipping the switch didn't accomplish anything.

In summary: if we were to accept *Invariance under Interpolated Variable Removal*, we will have to say that Fozzie Bear caused the money to be incinerated in *Coordination Game* iff Kermit the Frog caused the money to be incinerated in *Preemptive Overdetermination* (v_2). But it seems that Kermit's action was a cause and that Fozzie Bear's was not, and I see no plausible way of explaining away these appearances. So I have decided that we should reject *Invariance under Interpolated Variable Removal*. The theory from section 4 offers an explanation of why this principle is false: removing an interpolated variable from a path may remove a default consequence which lies along that path. So it may deprive us of an opportunity to stop tracing out consequences along that path.

Nonetheless, the theory from section 4 will never retract any of its verdicts as additional variables are interpolated along a path. That is: take a structural equations model \mathcal{M} which contains an interpolated variable $V \notin \text{CU}\{E\}$. Then, if you are able to trace a causal process from $(\mathbf{c}, \mathbf{c}^*)_{\mathbf{C}}$ to $(e, e^*)_E$ in \mathcal{M}^{-V} , you will still be able to trace a causal process from $(\mathbf{c}, \mathbf{c}^*)_{\mathbf{C}}$ to $(e, e^*)_E$ in \mathcal{M} .²⁴ Adding additional interpolated variables along a path may allow you to identify new

24. If the causal process in \mathcal{M}^{-V} did not involve V 's parent, Pa , or child, Ch —or if $Pa = E$ —then exactly the same process will be traceable in \mathcal{M} . Otherwise, there's some contrasts, pa^* and ch^* , such that $(pa, pa^*)_{Pa} \rightarrow (ch, ch^*)_{Ch}$ is a link in the causal process in \mathcal{M}^{-V} . Then, in \mathcal{M} , we can replace $(pa, pa^*)_{Pa} \rightarrow (ch, ch^*)_{Ch}$ with $(pa, pa^*)_{Pa} \rightarrow (f(pa), f(pa^*))_V \rightarrow (ch, ch^*)_{Ch}$, where f is the function from V 's structural equation, $V := f(Pa)$. Because V is interpolated, we don't have to worry about whether $(f(pa), f(pa^*))_V$ is default. Even if it is, it is immediately resolved with the consequence $(ch, ch^*)_{Ch}$.

causes which you couldn't have identified before. But it will never stop you from identifying causes which you could have identified before. Also, if $U \notin \mathbf{C}$ is a variable without any parents in \mathcal{M} , then you will be able to trace a causal process from $(c, c^*)_C$ to $(e, e^*)_E$ in \mathcal{M}^{-U} iff you are able to trace a causal process from $(c, c^*)_C$ to $(e, e^*)_E$ in \mathcal{M} .²⁵ So adding or removing parentless variables will never stop you from identifying causes which you could have identified before.

Sartorio (2005, 2013, 2016) defends a principle she calls the 'Causes as Difference-Makers' principle. According to this principle, if an event c caused e , then, if c hadn't occurred, c 's absence wouldn't have caused e . And, if c 's absence caused e , then, if c had occurred, c wouldn't have caused e . Let me generalise this principle so that it applies, not just to binary variables like whether an event occurs, but to variables with arbitrarily many values:

Causes as Difference-Makers If $C = c$, rather than c^* , is a cause of $E = e$, rather than e^* , then, if C had been c^* , then $C = c^*$, rather than c , would not have been a cause of $E = e$, rather than e^* .

According to the theory from section 4, this principle is true. For, if the rules allow you to trace a causal process from $(c, c^*)_C$ to $(e, e^*)_E$, then, if C had taken on the value c^* , the rules would not have allowed you to trace a causal process from $(c^*, c)_C$ to $(e, e^*)_E$.²⁶

25. I define \mathcal{M}^{-U} as the model that you get by replacing every occurrence of the variable U with its actual value wherever it appears in any structural equation. (See Gallow, 2021.) You will be able to trace all and only the same causal processes in \mathcal{M} that you are able to trace in \mathcal{M}^{-U} , except for those which are initiated by U itself.
26. Suppose (for the purposes of deriving a contradiction) that there is a causal process from $(c, c^*)_C$ to $(e, e^*)_E$ and that, were C to be c^* , there would be a causal process from $(c^*, c)_C$ to $(e, e^*)_E$. At least one of $(c, c^*)_C$ and $(c^*, c)_C$ is default. Without loss of generality, suppose $(c, c^*)_C$ is default. Then, either (i) $(c, c^*)_C$ is not resolved before we get to the consequence $(e, e^*)_E$ or (ii) it is. If (i), then the causal process leading from $(c, c^*)_C$ to $(e, e^*)_E$ must be a dependence process. And that means that $E = e$, rather than e^* , counterfactually depends upon $C = c$, rather than c^* (see footnote 16). In that case, were C to be c^* , there could not be a causal process from $(c^*, c)_C$ to $(e, e^*)_E$ for the simple reason that E would not be e , were $C = c^*$. Contradiction. On the other hand, if (ii), then $(c, c^*)_C$ must be resolved at a consequence $(r, r^*)_R$, and we must be able to trace a causal process from $(r, r^*)_R$ to $(e, e^*)_E$. Since $(r, r^*)_R$ lies in the dependence process, R would be r^* , were $C = c^*$. So, if C were c^* , there would have to be a causal process from $(r^*, r)_R$ to $(e, e^*)_E$. At least one of $(r, r^*)_R$ and $(r^*, r)_R$ is default, which means that all of the foregoing reasoning can be reiterated. Without loss of generality, suppose $(r, r^*)_R$ is default. Then, either (i) $(r, r^*)_R$ is not resolved before we reach the consequence $(e, e^*)_E$ or (ii) it is. If (i), then $E = e$ counterfactually depends upon $R = r$, rather than r^* , and so there cannot be a causal process from $(r^*, r)_R$ to $(e, e^*)_E$ for the simple reason that $E \neq e$ when $R = r^*$. Contradiction. If (ii), the foregoing reasoning reiterates again. The contradiction can be delayed, but so long as the causal process from $(c, c^*)_C$ to $(e, e^*)_E$ is finite, it cannot be delayed forever. So it cannot be that both there is a causal process from $(c, c^*)_C$ to $(e, e^*)_E$ and, if C were c^* , there would be a causal process from $(c^*, c)_C$ to $(e, e^*)_E$.

Hitchcock (2007) defends a principle which we can call ‘the Principle of Sufficient Reason.’²⁷ According to this principle, there is a special circumstance in which causation turns out to be equivalent to counterfactual dependence. Roughly, these are the circumstances in which all of the deviancy ‘in between’ C and E comes from C itself. That is: for any deviancy you find in a variable ‘in between’ C and E , that deviancy has a ‘sufficient reason’ for existing which is attributable to some other variables ‘in between’ C and E —and, ultimately, attributable to C itself. To state the principle more carefully, let me introduce a few definitions. Firstly, the ‘causal network connecting C to E ’, \mathbf{N} , is just the set containing every variable lying on a directed path of influence from C to E (including C and E themselves). For any $V \in \mathbf{N}$, let V ’s \mathbf{N} -parents be all the variables in \mathbf{N} which are also parents of V . (C does not have any \mathbf{N} -parents, but every other variable in \mathbf{N} will have \mathbf{N} -parents.) Hitchcock says that \mathbf{N} is ‘self-contained’ iff, for every variable $V \in \mathbf{N}$ other than C , when all of V ’s \mathbf{N} -parents take on default values, and V ’s other parents take on their actual values, V takes on a default value. Then, Hitchcock’s principle says that, in a self-contained network, counterfactual dependence is both necessary and sufficient for causation.

Almost all of the cases Hitchcock discusses in his 2007 paper involve binary variables which have one deviant value and one default value. In that special case, his principle is a consequence of the theory from section 4. Even when variables have arbitrarily many values, so long as each variable has exactly one default value, there is another, nearby principle which is also a consequence of the theory from section 4:

Principle of Sufficient Reason If the causal network connecting C to E is self-contained, if every variable in the network has exactly one default value, and if either c or c^* is default, then $C = c$, rather than c^* , is a cause of $E = e$ if and only if $E = e$ counterfactually depends upon $C = c$, rather than c^* .

This principle doesn’t quite give us conditions in which counterfactual dependence is both necessary and sufficient for causation. Instead, it gives us conditions for when $E = e$ counterfactually depending upon $C = c$, *rather than* c^* , is a necessary and sufficient condition for $C = c$, *rather than* c^* , being a cause of $E = e$. The theory from section 4 tells us that this will be the case so long as

27. Hitchcock (2007) gives this principle the name ‘TC’, for ‘token causation’, but his explanation of the principle appeals to something he names ‘the principle of sufficient reason.’

(1) the network connecting C to E is self-contained, (2) every variable in the network has exactly one default value, and (3) either c or c^* is default.²⁸ (These conditions hold in all of the ‘self-contained’ networks discussed in Hitchcock, 2007.)

What was supposed to be special about a ‘self-contained’ network, \mathbf{N} , was that we could attribute all of the deviancy within \mathbf{N} to the other variables in \mathbf{N} —and, ultimately, back to C . Hitchcock captured this idea with the requirement that, for any $V \in \mathbf{N}$, if V ’s \mathbf{N} -parents take on default values, and its non- \mathbf{N} -parents take on their actual values, then V should take on a default value, too. But, in cases where there are arbitrarily many values and multiple grades of deviancy, we might want to capture the idea with a slightly different requirement. Let’s say that the network connecting C to E , \mathbf{N} , is *self-sustained* iff, for every variable $V \in \mathbf{N}$ other than C , if some of V ’s \mathbf{N} -parents take on *more* default values, and V ’s other parents take on their actual values, then either V ’s value will be unchanged, or V will take on a more default value, too.²⁹ Then, it turns out

28. There are two cases to consider: either (i) c is default, or (ii) c^* is default. In case (i), (c, c^*) is default, and every variable in the network connecting C to E , \mathbf{N} , takes on its one and only default value (because the network is self-contained). So rule #4 requires us to trace out every consequence of (c, c^*) until it is resolved into a single deviant consequence. But (c, c^*) cannot be resolved into a single deviant consequence within \mathbf{N} , because this would require some $R \in \mathbf{N}$ having a deviant value when C ’s value is default. So, in tracing out the causal process emanating from (c, c^*) , we must be tracing out a dependence process within \mathbf{N} . In case (ii), (c, c^*) is a deviant consequence. And we can show (by induction) that, for every variable $V \in \mathbf{N}$, either V ’s value is default, or else the rules require us to include a deviant consequence for V in the process. C clearly has this property (base case). Take any $V \in \mathbf{N} \setminus \{C\}$, and suppose that all of the variables closer to C than V have the property (inductive hypothesis). Then, all of V ’s \mathbf{N} -parents either take on a default value or else have a deviant consequence included in the process. Suppose $V = v$ and that v isn’t default. Then, when we trace forward from all of the consequences for V ’s parents which we’ve already included in the process, we will have set all of V ’s \mathbf{N} -parents to default values, and all of V ’s other parents to their actual values. So V ’s value will have to become a default value, $v^* \neq v$, and rule #3 will require us to include $(v, v^*)_V$ in the process. So either V ’s value is default or else the rules require us to include a deviant consequence for V in the process. Since every variable has exactly one default value, the only possible consequences are deviant consequences. And so, whether we are in case (i) or case (ii), when we trace out the causal process emanating from (c, c^*) , we must include every possible consequence within \mathbf{N} and we are therefore tracing out a dependence process within \mathbf{N} . As we learnt in footnote 16, there is a dependence process from $(c, c^*)_C$ to (e, e^*) (for some $e^* \neq e$) if and only if $E = e$ counterfactually depends upon $C = c$, rather than c^* . So, within a self-contained network in which every variable has exactly one default value, if either c or c^* is default, then $C = c$, rather than c^* , is a cause of $E = e$ if and only if $E = e$ counterfactually depends upon $C = c$, rather than c^* .

29. More carefully: for any variable $V \in \mathbf{N}$, let \mathbf{Q} be V ’s non- \mathbf{N} -parents, and let \mathbf{P} be V ’s \mathbf{N} -parents. Then, V has a structural equation of the form $V := f(\mathbf{Q}, \mathbf{P})$. Let $\mathbf{q}_@$ be the actual values of \mathbf{Q} , and let \mathbf{p} and \mathbf{p}^* be two assignments of values to the variables in \mathbf{P} such that no value in \mathbf{p}^* is more deviant than the ‘corresponding’ value in \mathbf{p} . Then, \mathbf{N} is *self-sustained* iff, for every $V \in \mathbf{N}$, except for C , $f(\mathbf{q}_@, \mathbf{p}^*)$ is either the same value as $f(\mathbf{q}_@, \mathbf{p})$ or else $f(\mathbf{q}_@, \mathbf{p}^*)$ is more default

that, according to the theory from section 4, within a self-sustained network, counterfactual dependence is both necessary and sufficient for causation. That is, the theory validates the following variant of Hitchcock's principle:³⁰

Principle of Sufficient Reason (v2) If the causal network connecting C to E is self-sustained, then $C = c$ is a cause of $E = e$ if and only if $E = e$ counterfactually depends upon $C = c$.

than $f(\mathbf{q}@, \mathbf{p})$.

30. First note that, within a self-sustained network, you will never be able to trace out a default consequence from a collection of deviant consequences. For every $V \in \mathbf{N} \setminus \{C\}$, making some of V 's \mathbf{N} -parents more default either won't change V 's value, in which case rule #0 won't allow you to include a consequence for V in the process, or else it will make V 's value more default, in which case the consequence for V will be deviant. Now, either (i) $(c, c^*)_C$ is deviant or (ii) it is default. If (i), then all of its consequences within \mathbf{N} will be deviant, and rule #3 will require you to trace all of them out. The same holds for all of the possible consequences of $(c, c^*)_C$ within \mathbf{N} , and all of their possible consequences within \mathbf{N} , and so on. Since they are all deviant, all of their possible consequences within \mathbf{N} are deviant, too. So, within \mathbf{N} , rule #3 will require you to trace out a dependence process. On the other hand, if (ii), then rule #4 will require you to trace out every possible consequence of $(c, c^*)_C$ until $(c, c^*)_C$ is resolved into a deviant consequence. If $(c, c^*)_C$ doesn't resolve into a deviant consequence within \mathbf{N} , then, in tracing the causal process from $(c, c^*)_C$, you will be tracing out a dependence process within \mathbf{N} . If, on the other hand, $(c, c^*)_C$ does resolve into a deviant consequence, $(r, r^*)_R$, for some $R \in \mathbf{N}$, then every possible consequence of $(r, r^*)_R$ will be deviant (since \mathbf{N} is self-sustained), and rule #3 will require that all of these consequences be included. So, no matter whether (i) or (ii), in tracing out the causal process from $(c, c^*)_C$, you will be tracing out a dependence process within \mathbf{N} . As we saw in footnote 16, it is possible to trace out a dependence process from $(c, c^*)_C$ to $(e, e^*)_E$ if and only if $E = e$, rather than e^* , counterfactually depends upon $C = c$, rather than c^* . So, within a self-sustained network, counterfactual dependence is both necessary and sufficient for causation.

References

- Andreas, Holger & Günther, Mario. 2020. "Causation in Terms of Production." In *Philosophical Studies*, 177: 1565–1591. <https://doi.org/10.1007/s11098-019-01275-3>. [20]
- Andreas, Holger & Günther, Mario. 2021. "A Ramsey Test Analysis of Causation for Causal Models." In *The British Journal for the Philosophy of Science*, 72 (2): 587–615. [20]
- Andreas, Holger & Günther, Mario. forthcoming. "Difference-Making Causation." In *The Journal of Philosophy*. [20]
- Aronson, Jerrold. 1971. "On the Grammar of 'Cause'." In *Synthese*, 22 (3/4): 414–430. [20]
- Beckers, Sander & Vennekens, Joost. 2017. "The Transitivity and Asymmetry of Actual Causation." In *Ergo*, 4: 1–27. [20]
- Beckers, Sander & Vennekens, Joost. 2018. "A Principled Approach to Defining Actual Causation." In *Synthese*, 195 (2): 835–862. [20]
- Blanchard, Thomas & Schaffer, Jonathan. 2017. "Cause without Default." In *Making a Difference*, edited by Helen Beebe, Christopher Hitchcock, & Huw Price, Oxford: Oxford University Press. [6], [8]
- Bochman, Alexander. 2021. *A Logical Theory of Causality*. Cambridge, MA: The MIT Press. [20]
- Briggs, R. A. 2012. "Interventionist Counterfactuals." In *Philosophical Studies*, 160: 139–166. [11]
- Collins, J. 2004. "Preemptive Prevention." In Collins *et al.* (2004), chapter 4, 107–117. [19], [20]
- Collins, J., Hall, N., & Paul, L. A. (editors). 2004. *Causation and Counterfactuals*. Cambridge, MA: The MIT Press. [29], [30], [32], [34]
- Dowe, Phil. 2000. *Physical Causation*. Cambridge: Cambridge University Press. [20]
- Eells, Ellery. 1991. *Probabilistic Causality*. Cambridge: Cambridge University Press. [20]

- Ehring, Douglas. 1997. *Causation and Persistence*. Oxford: Oxford University Press. [20]
- Fair, David. 1979. "Causation and the Flow of Energy." In *Erkenntnis*, 14: 219–50. [20]
- Galles, David & Pearl, Judea. 1998. "An axiomatic characterization of causal counterfactuals." In *Foundations of Science*, 3 (1): 151–182. [11]
- Gallow, J. Dmitri. 2016. "A Theory of Structural Determination." In *Philosophical Studies*, 173 (1): 159–186. [5]
- Gallow, J. Dmitri. 2021. "A Model-Invariant Theory of Causation." In *Philosophical Review*, 130 (1): 45–96. [6], [8], [10], [20], [21], [25]
- Gallow, J. Dmitri. ms. "Causal Counterfactuals without Backtracking or Miracles." [5]
- Gerstenberg, Tobias & Icard, Thomas. 2020. "Expectations Affect Physical Causation Judgements." In *Journal of Experimental Psychology*, 149 (3): 599–607. [8]
- Hall, Ned. 2000. "Causation and the Price of Transitivity." In *Journal of Philosophy*, 97 (4): 198–222. Reprinted in Collins *et al.* (2004). [7]
- Hall, Ned. 2004. "Two Concepts of Causation." In Collins *et al.* (2004), 225–276. [5], [15], [18]
- Hall, Ned. 2007. "Structural Equations and Causation." In *Philosophical Studies*, 132 (1): 109–136. [5], [8], [9], [20]
- Halpern, Joseph Y. 2008. "Defaults and Normality in Causal Structures." In *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning*, 198–208. [8]
- Halpern, Joseph Y. 2016. *Actual Causality*. Cambridge, MA: MIT Press. [8], [20]
- Halpern, Joseph Y. & Hitchcock, Christopher. 2015. "Graded Causation and Defaults." In *The British Journal for the Philosophy of Science*, 66 (2): 413–457. [8]
- Halpern, Joseph Y. & Pearl, Judea. 2001. "Causes and Explanations: A Structural-Model Approach. Part 1: Causes." In *Proceedings of the Seventeenth*

- Conference on Uncertainty in Artificial Intelligence*, edited by John Breese & Daphne Koller. San Francisco: Morgan Kaufman, 194–202. [20]
- Halpern, Joseph Y. & Pearl, Judea. 2005. “Causes and Explanations: A Structural-Model Approach. Part 1: Causes.” In *The British Journal for the Philosophy of Science*, **56**: 843–887. [7], [20]
- Hart, H.L.A. & Honoré, Tony. 1985. *Causation in the Law*. Oxford: Clarendon Press, second edition. [8]
- Hiddleston, Eric. 2005a. “A Causal Theory of Counterfactuals.” In *Noûs*, **39** (4): 632–657. [11]
- Hiddleston, Eric. 2005b. “Causal Powers.” In *The British Journal for the Philosophy of Science*, **56**: 27–59. [5]
- Hitchcock, Christopher. 1996a. “Farewell to Binary Causation.” In *Canadian Journal of Philosophy*, **26** (2): 267–282. [10]
- Hitchcock, Christopher. 1996b. “The Role of Contrast in Causal and Explanatory Claims.” In *Synthese*, **107** (3): 395–419. [10]
- Hitchcock, Christopher. 2001. “The Intransitivity of Causation Revealed in Equations and Graphs.” In *The Journal of Philosophy*, **98** (6): 273–299. [5], [20]
- Hitchcock, Christopher. 2007. “Prevention, Preemption, and the Principle of Sufficient Reason.” In *Philosophical Review*, **116** (4): 495–532. [1], [8], [9], [20], [25], [26], [27]
- Hitchcock, Christopher. 2009. “Structural Equations and Causation: Six Counterexamples.” In *Philosophical Studies*, **144**: 391–401. [20]
- Hitchcock, Christopher. 2011. “Trumping and contrastive causation.” In *Synthese*, **181**: 227–240. [10]
- Hitchcock, Christopher & Knobe, Joshua. 2009. “Cause and Norm.” In *Journal of Philosophy*, **106** (11): 587–612. [8]
- Icard, Thomas F., Kominsky, Jonathan F., & Knobe, Joshua. 2017. “Normality and Actual Causal Strength.” In *Cognition*, **161**: 80–93. [8]
- Kahneman, Daniel & Miller, Dale T. 1986. “Norm Theory: Comparing Reality to Its Alternatives.” In *Psychological Review*, **94** (2): 136–153. [8]

- Lewis, David K. 1973. "Causation." In *The Journal of Philosophy*, 70 (17): 556–567. [20]
- Lewis, David K. 1986. "Postscripts to 'Causation.'" In *Philosophical Papers*, New York: Oxford University Press, volume II. [20]
- Lewis, David K. 2004. "Causation as Influence." In Collins *et al.* (2004), chapter 3, 75–106. [20]
- Mackie, John L. 1965. "Causes and Conditions." In *American Philosophical Quarterly*, 2 (4): 245–55. [20]
- Maslen, Cei. 2004. "Causes, contrasts, and the nontransitivity of causation." In Collins *et al.* (2004), 341–357. [10], [16]
- Maudlin, Tim. 2004. "Causation, Counterfactuals, and the Third Factor." In Collins *et al.* (2004), 419–443. [8]
- McDermott, Michael. 1995. "Redundant Causation." In *The British Journal for the Philosophy of Science*, 46 (4): 523–544. [19], [20]
- McGrath, Sarah. 2005. "Causation by Omission: A Dilemma." In *Philosophical Studies*, 123: 125–148. [8]
- Menzies, Peter. 2004. "Causal Models, Token Causation, and Processes." In *Philosophy of Science*, 71 (5): 820–832. [8]
- Menzies, Peter. 2007. "Causation in Context." In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by Huw Price & Richard Corry, Oxford: Clarendon Press, chapter 8, 191–223. [8]
- Menzies, Peter. 2017. "The Problem of Counterfactual Isomorphs." In *Making a Difference: Essays on the Philosophy of Causation*, edited by Helen Beebe, Christopher Hitchcock, & Huw Price, Oxford: Oxford University Press. [8]
- Paul, L. A. & Hall, Ned. 2013. *Causation: A User's Guide*. Oxford: Oxford University Press. [8]
- Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press, second edition. [7]
- Ramachandran, Murali. 1997. "A Counterfactual Analysis of Causation." In *Mind*, 106 (422): 263–277. [20]

- Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press. [20]
- Salmon, Wesley. 1994. "Causality without Counterfactuals." In *Philosophy of Science*, **61** (2): 297–312. [20]
- Sartorio, Carolina. 2005. "Causes as Difference-Makers." In *Philosophical Studies*, **123**: 71–98. [7], [25]
- Sartorio, Carolina. 2013. "Making a Difference in a Deterministic World." In *Philosophical Review*, **122** (2): 189–214. [25]
- Sartorio, Carolina. 2016. *Causation & Free Will*. Oxford: Oxford University Press. [1], [7], [25]
- Schaffer, Jonathan. 2000. "Causation by Disconnection." In *Philosophy of Science*, **67** (2): 285–300. [18]
- Schaffer, Jonathan. 2001. "Causes as Probability Raisers of Processes." In *Journal of Philosophy*, **98**: 75–92. [20]
- Schaffer, Jonathan. 2005. "Contrastive Causation." In *The Philosophical Review*, **114** (3): 297–328. [10], [16]
- Schaffer, Jonathan. 2012a. "Causal Contextualism." In *Contrastivism in Philosophy*, edited by Blaauw, Routledge, chapter 2, 35–63. [10]
- Schaffer, Jonathan. 2012b. "Disconnection and Responsibility." In *Legal Theory*, **18** (4): 399–435. [18]
- Suppes, Patrick. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing Company. [20]
- Thomson, Judith Jarvis. 2003. "Causation: Omissions." In *Philosophy and Phenomenological Research*, **66** (1): 81–103. [8]
- Weslake, Brad. forthcoming. "A Partial Theory of Actual Causation." In *The British Journal for the Philosophy of Science*. [20]
- Wolff, J. E. 2016. "Using Defaults to Understand Token Causation." In *The Journal of Philosophy*, **113** (1): 5–26. [8]
- Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press. [20]

Wysocki, Tom. ms. "Conjoined Cases." [8]

Yablo, Stephen. 2004. "Advertisement for a Sketch of an Outline of a Prototheory of Causation." In Collins *et al.* (2004), chapter 5, 119–138. [20]