Dependence, Defaults, and Needs

J. DMITRI GALLOW

1 | INTRODUCTION

Fozzie Bear and Crazy Harry play a coordination game. Both players have switches in front of them with two positions: Left and Right. Between them sits $1,000,000. Fozzie has first move. He can either flip his switch or leave it as it is. Next, after learning what Fozzie has done, Harry can either flip his switch or leave it alone. If the switches are in the same position, then the money will be distributed between the players. If the switches are in different positions, the money will be incinerated in an extravagant explosion. To start, Fozzie's switch is set to Left, and Harry's is set to Right. The game begins. Wanting the money, Fozzie flips his switch to Right. Seeing this, and wanting the explosion, Crazy Harry flips his switch to Left. The money is incinerated.[1]

Many of us are inclined to say that Fozzie Bear's flipping his switch to Right did not cause the money to be incinerated. Asked to explain why, it's quite natural to say something like this: "Crazy Harry just wanted the explosion. So, had Fozzie left his switch set to Left, Harry wouldn't have flipped his switch to Left, and the money would have been incinerated anyhow. So the money was going to be incinerated whether Fozzie flipped or not—it was *on a course* to be incinerated either way—and Fozzie's flipping the switch didn't make any difference." This kind of response is incredibly natural, but the study of causation has taught many of us to repress it. We've learnt to equate 'making a difference' with counterfactual dependence, and 'being on a course to happen either way' with the lack of counterfactual dependence. And we've learnt that counterfactual dependence is not necessary for causation.

To see why dependence isn't necessary for causation, suppose that Waldorf has also rigged the money with explosives, and he and Statler watch the coordination game while cracking jokes from the balcony. Had Crazy Harry not flipped his switch to Left, Waldorf would have detonated his explosives, and the money would still have been incinerated. In that case, the incineration wouldn't depend upon Harry's flipping his switch to Left any more than it depends upon Fozzie Bear's flipping his switch to Right. Even so, many of us share the judgement that, unlike Fozzie, Harry *did* cause the money to be incinerated. Waldorf's explosives were a *backup* for the incineration; but that backup was *preempted* by Crazy Harry's flipping his switch to Left. The backup would have been a cause, were it not preempted, but is not in fact a cause. In any case like this, there will be causation without counterfactual dependence.

Draft of February 25, 2021; Word Count: 9,762
✉: dmitri.gallow@acu.edu.au

1. The case is modelled on McDermott (1995)'s *Shock C*.

Lewis (2004) concludes that we never had a good reason to think that Fozzie Bear didn't cause the money to be incinerated. He challenges those who think otherwise: "if you ever accept preemptive causation, you must have learned to resist [the idea that causation requires counterfactual dependence]. Why yield to it now?" (p. 99). If we grant Lewis that these thoughts about Fozzie 'not making a difference' should be understood in terms of the incineration not depending upon his action, then he is correct. Our reason for not holding Fozzie causally responsible applies just as well to Crazy Harry, who clearly caused the money to be incinerated. But it's noteworthy that, although there is a strong temptation to say that Fozzie didn't make a difference to whether the money was incinerated, there isn't a corresponding temptation to say that Harry didn't make a difference. The temptation doesn't arise; and, when forced to consider the question, the thought I find most natural is: 'If *anyone* made a difference, it was Crazy Harry'.

Here, I'm going to suggest an alternative way of understanding these thoughts about 'not making a difference' and 'being on a course to happen anyhow'. It builds on the following rough and programmatic idea: while the money's incineration *already had* all it needed to occur without Fozzie Bear's action, the same cannot be said for Crazy Harry's. Without Harry's flipping the switch to Left, the money's incineration would have needed something *additional* from Waldorf in order to happen. When we're tempted by the thought that Fozzie *didn't make a difference*, what's true in this thought is that Fozzie didn't contribute anything which was needed—the incineration already had all it needed without his action. In contrast, Harry did contribute something which was needed—without Harry's contribution, the incineration did not have all it needed to happen. (It is of course true that, had Harry not given the incineration what it needed to happen, Waldorf would have given it instead. Even so, Harry gave what was needed, and Waldorf did not.)

In §2, I will say a bit more to develop this puzzle; I'll show that cases like these give rise to a trilemma. I will opt for one horn of this trilemma, which leads me in §3 to precisify the rough and programmatic thoughts above into a theory of when some occurrence *already had all that it needed to happen*, and when it did not. In §4, I'll explain how this notion can be incorporated into a theory of causation I've developed elsewhere (Gallow, forthcoming) and how it allows us to distinguish Fozzie Bear from Crazy Harry. In the appendix, I'll provide a careful statement of the resulting theory.

2 | PREEMPTERS AND SELF-UNDERMINERS

When he flipped his switch to the Left, Crazy Harry did two things. Firstly, he caused the money to be incinerated. Secondly, he kept Waldorf from incinerating the money. Waldorf was a *backup*, potential cause of the money's incineration. And this backup was *preempted* by Crazy Harry's actions. Let's call this case '*Preempter*'.

To explore the structure of *Preempter*, I'll introduce three binary variables, H, W, and I. H is a variable which takes the value 1 if Harry flips his switch, and takes the value 0 if he does not. Likewise, $W = 1$ stands for Waldorf igniting his explosives, and $W = 0$ stands for Waldorf doing nothing. Finally $I = 1$ represents the money being incinerated, while $I = 0$ stands for the nothing happening to the money. We can encode the counterfactual structure between these three variables with this system of

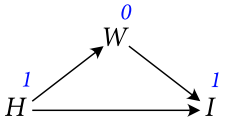
equations:

$$H = 1$$
$$W := \neg H$$
$$I := H \vee W$$

Here, '$\neg$' and '$\vee$' are the familiar functions 'not' and 'or'. The equation '$H = 1$' tells us that Harry in fact flips his switch. The *structural* equation $W := \neg H$ provides the following counterfactual information: if Harry were to flip, then Waldorf would not ignite his explosives, and if Harry were to not flip, then Waldorf would ignite his explosives. And the final equation, '$I := H \vee W$', tells us that, if either Harry were to flip or Waldorf were to ignite his explosives, the money would be incinerated. Notice that there's an asymmetry between the left- and right-hand-sides of a structural equation. That's why I write ':=' instead of '='. We can solve for the value of every variable in this system of equations by just starting at the top and working our way down. We know that $H = 1$, so the second equation tells us that $W = \neg 1 = 0$. And if we know that $H = 1$ and $W = 0$, then the final equation tells us that $I = 1$.

When placed together in a single system of equations like this, the pair of structural equations $W := \neg H$ and $I := H \vee W$ tell us more than either does on its own. For instance, from this system of equations, we can deduce that, were Harry to not flip (were H to take on the value 0) the money would still be incinerated (I would still equal 1). But neither the equation $W := \neg H$ nor the equation $I := H \vee W$ individually will tell us this. Systems of structural equations like these lie at the heart of contemporary causal modelling approaches to causation. They are discussed in more depth elsewhere,[2] but for our purposes, we need only understand that the equations encode everything there is to know about the counterfactual relationship between the variables H, W, and I, as well as their actual values.

When I'm thinking about a system of equations, I find it helpful to draw a directed graph like this:



This graph tells you that H shows up on the right-hand-side of W's structural equation, and both H and W show up on the right-hand-side of I's structural equation. We can use the metaphor of genealogy to talk about the relationship between variables in a graph like this. For instance, H and W are I's *causal parents*, and W is one of H's *causal children*. I've also decorated the graph with variables' actual values; so the graph tells us that $H = 1$, $W = 0$, and $I = 1$.

Let me briefly say something about why I'm modelling *Preempter* with systems of equations like these. I am drawn towards a counterfactual theory of causation, accord-

2. I've said more about how I think we should understand systems of equations like these in Gallow (2016, ms). See also, for instance, Hitchcock (2001, 2007), Woodward (2003), Hall (2007), Pearl (2000), Halpern & Pearl (2005), and Paul & Hall (2013).

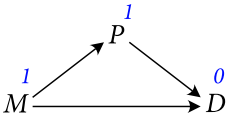
ing to which causation is revealed through counterfactual dependence. The simplest possible counterfactual theory *identifies* causation with dependence, but cases like *Preempter* show us that this theory cannot be correct. Harry's flipping his switch caused the money to be incinerated, but the incineration does not depend upon Harry's flip. Starting with Lewis (1973), the hope was that, even though causation cannot be identified with dependence, it *can* be identified with some other kind of counterfactual structure. For instance, Lewis thought that causation should be identified with the ancestral, or the transitive closure, of dependence. Like many, I've come to think that Lewis's view is incorrect, for various and sundry reasons. For one, consider:

> *Thwarted Assassination*
> Moriarty recites an obscure line of poetry. This is a coded signal ordering his henchman to poison Holmes's drink. Holmes, having already worked out the code, is alerted by the poem, and empties his glass into a nearby house plant.

Holmes's survival depends upon him not drinking. Had he drank, he would have died. And his not drinking depends upon Moriarty's poetry recital. Had Moriarty not recited that line of poetry, Holmes would have drank. So there is a chain of dependence leading from Moriarty's recitation to Holmes's survival. Lewis is forced to conclude that Moriarty's coded order to poison Holmes's drink caused Holmes to survive. This seems to me to be the wrong result. I am inclined to say that Holmes survived *in spite of* Moriarty's poetic assassination attempt—not *because* of it. (Though we should be careful. It's true that Moriarty's giving the order *in that code*, rather than some other way which would have escaped Holmes's attention, caused Holmes to survive. What's false is that Moriarty's giving the order to assassinate Holmes *at all*, rather than giving no order, is a cause of Holmes's survival.[3])

As I said above, the hope *was* that causation could be identified with some richer counterfactual structure. But by now, I think that hope should have died. For the work of Hiddleston (2005) and Hall (2007) has taught us that *no* amount of counterfactual information on its own will be sufficient to determine whether or not $C = c$ caused $E = e$ (for any variables, C and E, with values c and e, respectively). To appreciate their point, let's write up the natural system of structural equations for *Thwarted Assassination*. Let $M = 1$ stand for Moriarty giving the coded order to poison Holmes, and let $M = 0$ represent Moriarty not giving any order. Let $P = 1$ represent the henchman poisoning Holmes's drink, and let $P = 0$ represent the henchman not poisoning the drink. Finally, let $D = 1$ represent Holmes dying, and let $D = 0$ represent him not dying. Then, the following system of equations encodes all of the relevant counterfactual structure.

$$\begin{aligned} M &= 1 \\ P &:= M \\ D &:= P \wedge \neg M \end{aligned}$$

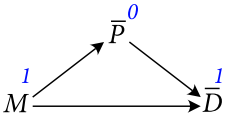


(Here, '$\wedge$' is 'and'.) Moriarty gives the order. Whether the henchman poisons the drink depends upon whether Moriarty gives the order. And whether Holmes dies depends

3. For more on the role of contrasts in causation, see Hitchcock (1996a,b), Maslen (2004), Schaffer (2005), and Hitchcock (2011).

upon whether Moriarty gives the order and whether the drink is poisoned. The only way Holmes will die is if the drink gets poisoned without Moriarty's order alerting Holmes to the danger. HIDDLESTON and HALL's observation is that the counterfactual information encoded in this system of equations is in fact *isomorphic* to the information encoded in the system of equations for *Preempter*. This isn't obvious, but we can more clearly appreciate the isomorphism if we exchange the variables P and D for the variables $\overline{P} \stackrel{\text{def}}{=} 1 - P$ and $\overline{D} \stackrel{\text{def}}{=} 1 - D$. In other words, $\overline{P}$ takes on the value 1 if the henchman does *not* poison the drink, and it takes on the value 0 if he *does*. And $\overline{D}$ takes the value 1 if Holmes does *not* die, and it takes the value 0 if he *does*. With these variables, the counterfactual structure of the case is represented with this system of equations:

$$\begin{aligned} M &= 1 \\ \overline{P} &:= \neg M \\ \overline{D} &:= \overline{P} \vee M \end{aligned}$$



And these equations are precisely the same as the equations we wrote down for the case of *Preempter*, with 'M' exchanged for 'H', '$\overline{P}$' exchanged for 'W', and '$\overline{D}$' exchanged for 'I'.

I conclude that no amount of information about the wider counterfactual structures in which C and E are ensconced is enough on its own to determine whether c caused e. (A quick word on notation: throughout, I will understand schematic expressions like 'c caused e' as shorthand for '$C = c$ caused $E = e$', for some variables C and E with values c and e, respectively. Likewise, when I say things like 'c happens', I understand this to mean that $C = c$.) If we are to distinguish cases like *Preempter* from cases like *Thwarted Assassination*, we will have to look to some other kind of information about the cases. And it is noteworthy that, while the variable value $W = 0$ stands for the *default* state of Waldorf doing nothing, $\overline{P} = 0$ stands for a *departure* from the default state of the henchman doing nothing. Likewise, while $I = 1$ stands for the event of the money being incinerated—a departure from the default state of the money remaining as it is—$\overline{D} = 0$ stands for the default state of Holmes continuing to survive. For this and other reasons, several authors have suggested that, in order to give an adequate theory of causation, we will need information about which variable values represent *default* states, and which represent departures therefrom.[4] Just to have a term, if a variable value is a deviation from the default, then we can say that that value is *deviant*.

I'm afraid I won't have the space to say very much about this difference between variable values which are *default* those 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 a default value. On the other hand, if it feels natural to describe a variable value by saying that something *happened*, or that it rep-

4. See, in particular, KAHNEMAN & MILLER (1986), THOMSON (2003), MCGRATH (2005), HALL (2007), HITCHCOCK (2007), HALPERN (2008), HITCHCOCK & KNOBE (2009), HALPERN (2016), PAUL & HALL (2013), and HALPERN & HITCHCOCK (2015).

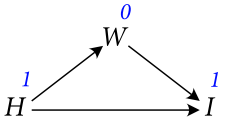
resents an *event*, then that variable value is likely a deviant departure from the default. Another helpful characterisation of default states can be given in terms of what we're inclined to imagine when we counterfactually suppose that some event did not take place. When we counterfactually suppose that the henchman didn't poison Holmes's drink, we're not inclined to imagine the henchman attacking him with a dagger, or shooting him with a revolver, or staging a production of *West Side Story*. Instead, we simply imagine him standing there, *doing nothing*, or perhaps making a drink 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 one of the *default* values of that variable. 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 henchman didn't poison the drink, we easily imagine him just standing about. But, if the henchman is just standing about, and we're asked to counterfactually suppose that he *isn't* just standing about, we're unsure what we're being asked to imagine, unless some salient contrast has already been suggested.[5]

So, we when are modelling a causal scenario with a system of structural equations, we will have to additionally specify which values of a variable are default, and which are deviant. So long as we're confining attention to binary variables (variables with only two potential values), we can encode this information by adopting the convention of using '0' for the default value, and '1' for the deviant value. Thus, with our model of *Preempter*, $H = 1$ stands for the deviant event of Harry flipping his switch, and $H = 0$ stands for the default state of Harry doing nothing. Likewise, $W = 1$ stands for the deviant event of Waldorf igniting his explosives, and $W = 0$ stands for the default state of him doing nothing.

Let's return to the question 'What's the difference between Fozzie Bear and Crazy Harry?'. In the remainder of this section, I will crystallise our nascent concerns from the introduction into a trilemma. Let me begin by slightly modifying *Preempter*. Let's suppose that, while either Harry's flipping his switch or Waldorf's igniting his explosives would be enough to incinerate the money *on their own*, had Harry flipped his switch *and* Waldorf ignited his explosives, the money would *not* have been incinerated. We may imagine that Waldorf's ignition button is plugged into the same electrical outlet as Fozzie and Harry's switches, so that, if the switches are misaligned and Waldorf pushes his button, then there will be a power surge, and there will not be any explosion.

I don't think this modification affects whether Harry's flipping the switch caused the money to be incinerated. Even with this change in place, Harry caused the money to be incinerated. We can model this modified version of *Preempter* with this causal model, which I will call '$\mathcal{M}_P$':

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



In I's equation, I am using $[\phi]$ to stand for the truth-value of ϕ. So $[H \neq W]$ is the

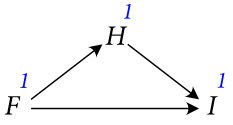
5. Cf. Hall (2007, §6)

exclusive disjunction of H and W. It is 1 iff *exactly* one of H and W is 1, and the other is 0. Otherwise, $[H \neq W]$ will be 0.

Let's now consider the relationship between Fozzie Bear, Harry, and the explosion. Let's focus on a version of the case where Waldorf is absent. So suppose that there are no backup explosives, and that the money will be incinerated if, and *only* if, Fozzie and Harry's swiches are misaligned. In this case, Fozzie flips his switch in an attempt to prevent the money from being incinerated, but this attempt ends up undermining itself by leading Harry to flip his switch, which leads to the money's incineration. So I'll call this case *Self-Underminer*. (To be clear, *Self-Underminer* and *Preempter* are two different cases, even though they share some of the same characters.)

To model *Self-Underminer*, let $F = 1$ stand for Fozzie flipping his switch, and let $F = 0$ stand for Fozzie doing nothing. H and I are as the were in $\mathcal{M}_P$. Then, the counterfactual structure of *Self-Underminer* can be encoded in this model, which I'll call '$\mathcal{M}_S$':

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



If we're going to theorise about causation in terms of causal models like these, then we will want a theory which can tell us, whenever we hand it a model containing the variables C and E, whether c caused e or not. A theory like this will return verdicts, not about whether c caused e *full stop*, but rather about whether c caused e *in some causal model*. In Gallow (forthcoming, ms), I contend that a theory of causation like this should be *model-invariant*, in the sense that its verdicts shouldn't change if the model is superficially changed by including or excluding inessential variables. More carefully, I impose the following constraint on a theory of causation:

Interpolated Variable Removal If a variable V is interpolated along a path in a causal model $\mathcal{M}$, then $C = c$ is a cause of $E = e$ in $\mathcal{M}$ if and only if $C = c$ is a cause of $E = e$ in $\mathcal{M}^{-V}$.[6]

This principle requires some explanation. I say that a variable V is *interpolated along a path* iff it has a single causal parent, Pa, a single causal child, Ch,

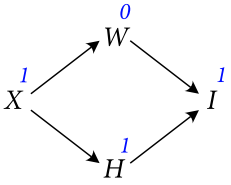
$$Pa \to V \to Ch$$

and Pa is not *also* a parent of Ch. If V is interpolated in the model $\mathcal{M}$, then $\mathcal{M}$ will include a structural equation of the form $V := f(Pa)$, for some function f of the variable Pa, and a structural equation $Ch := g(...V...)$, 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 an interpolated variable, then, to get the model $\mathcal{M}^{-V}$, you just get rid of the equation $V := f(Pa)$ entirely, and replace the equation $Ch := g(...V...)$ with $Ch := g(...f(Pa)...)$. The principle tells us that whether you include an interpolated 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.

6. Assuming, of course, that $V \neq C, E$.

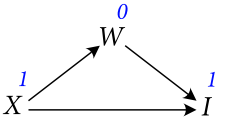
Now, consider the causal model shown below, which I'll call '$\mathcal{M}_X$'.

$$\begin{aligned} X &= 1 \\ H &:= X \\ W &:= \neg X \\ I &:= [H \neq W] \end{aligned}$$



(The variables are all binary, and the value 1 is more deviant than the value 0; beyond this, don't worry about how to interpret them.) In this model, W and H are both interpolated. They both have the sole parent X, and the sole child I. And X is not also a parent of I. If we remove H, we get the model $\mathcal{M}_X^{-H}$,

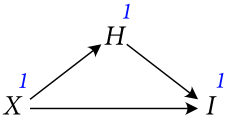
$$\begin{aligned} X &= 1 \\ W &:= \neg X \\ I &:= [X \neq W] \end{aligned}$$



To get the equation $I := [X \neq W]$, we just substituted the right-hand-side of H's equation, X, in for H in the equation $I := [H \neq W]$. Notice that the model $\mathcal{M}_X^{-H}$ is isomorphic to the model $\mathcal{M}_P$ of *Preempter*. Swap the variable 'X' out for 'H', and the models are precisely the same. By the isomorphism, any theory will say that $H = 1$ is a cause of $I = 1$ in $\mathcal{M}_P$ iff it says that $X = 1$ is a cause of $I = 1$ in $\mathcal{M}_X^{-H}$. Since Crazy Harry is a cause of the money's incineration, we want our theory to tell us that $H = 1$ is a cause of $I = 1$ in $\mathcal{M}_P$. So the theory must also tell us that $X = 1$ is a cause of $I = 1$ in $\mathcal{M}_X^{-H}$. Then, according to the principle **Interpolated Variable Removal**, $X = 1$ must count as a cause of $I = 1$ in $\mathcal{M}_X$, as well.

On the other hand, if we remove the interpolated variable W from $\mathcal{M}_X$, we get the model $\mathcal{M}_X^{-W}$,

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



To get the equation $I := [X = H]$, we just substituted the right-hand-side of W's equation, $\neg X$, in for W in the equation $I := [H \neq W]$, to get $I := [H \neq \neg X]$. But of course, H will not be equal to $\neg X$ iff H is equal to X, so this equation is equivalent to $I := [X = H]$. Notice that the model $\mathcal{M}_X^{-W}$ is isomorphic to the model $\mathcal{M}_S$ of *Self-Underminer*. Swap out 'X' for 'F', and the models are precisely the same. So $F = 1$ will count as a cause of $I = 1$ in $\mathcal{M}_S$ iff $X = 1$ counts as a cause of $I = 1$ in $\mathcal{M}_X^{-W}$. Since Fozzie Bear is *not* a cause of the money's incineration, we want our theory to tell us that $X = 1$ is *not* a cause of $I = 1$ in $\mathcal{M}_X^{-W}$. Then, according to the principle **Interpolated Variable Removal**, $X = 1$ must not count as a cause of $I = 1$ in $\mathcal{M}_X$, either.

But now we've contradicted ourselves. We've reasoned our way to the conclusion that $X = 1$ both is and isn't a cause of $I = 1$ in the model $\mathcal{M}_X$. Moreover, it doesn't look like appealing to information about which variable values are default and which are deviant will do anything to block the reasoning above. For, in both the model $\mathcal{M}_P$ and the model $\mathcal{M}_S$, we used '0' for events which are default, and '1' for events which

are deviant. Moreover, the values of all the relevant variables stand for *precisely the same things*. Both $F = 0$ and $H = 0$ stand for Fozzie Bear and Crazy Harry doing nothing, and both $F = 1$ and $H = 1$ stand for them flipping their switches. $I = 1$ stands for the money being incinerated, and $I = 0$ stands for nothing happening to the money.

There's some reason to think that, in general, actions which are moral or normatively expected are more default than actions which are immoral or normatively unexpected.[7] So we might think that, since we normatively *expected* Fozzie Bear to flip his switch, $F = 1$ should count as the default; whereas, since we normatively expect Crazy Harry to *refrain* from flipping his (though, of course, we don't *descriptively* expect this), $H = 0$ should count as a default. For a while, I was tempted by this response, but I've come to think it's a misdiagnosis. While it may help us get the right prediction in this case, it gets the wrong prediction in nearby variants. Suppose that the fireworks will go off iff the switches are misaligned. In an attempt to ruin everybody's fun, Fozzie flips one of the switches to Right. Fortunately, somebody installed a safeguard system, which will flip the other switch to Left if Fozzie's switch happens to be set to Right; so, after Fozzie flips his switch to Right, the other switch is flipped to Left, and the fireworks are lit. In this case, Fozzie's action is not normatively expected, but neither is it a cause of the fireworks display. In response to this kind of case, we might suggest that, when a fireworks display is scheduled, its occurrence counts as default. But we don't need to assume that the fireworks were scheduled, and assuming that they were unplanned doesn't seem to interfere with the judgement that Fozzie wasn't responsible for them. Moreover, if we assume that scheduled fireworks displays are always default, then it's easy to come up with isomorphisms showing that a causally efficacious planned fireworks display must be treated like a causally inefficacious 'bogus preventer'—I'll leave the details in this footnote for the curious reader.[8] We could take this dialectic further, but at the end of the day, I just don't see how to construct a genuinely predictive theory around this idea.

As I see it, there are three ways we might try to avoid the contradiction. In the first place, we might deny one of the judgements about *Preempter* and *Self-Underminer*. Secondly, we could deny the principle **Interpolated Variable Removal**. And thirdly, we could say that a causal model (understood to include information about which variable values represent default states and which represent deviations therefrom) does not tell us enough to determine which variable values are causally related and which are not.

7. See, for instance, McGrath (2005) and Hitchcock & Knobe (2009).

8. First case (*bogus prevention*): Holmes's drink is not poisoned, but Watson adds an antidote which would neutralise any poison, were the drink poisoned. In *bogus prevention*, we have the equation $D := P \wedge \neg A$ (that is: Holmes will *d*ie iff there's *p*oison and not any *a*ntidote). For each variable, 1 is deviant and 0 is default. Second case: there are two fireworks displays planned, one on the northside and one on the southside. On the northside, the fireworks are set to go off at 9:00. On the southside, they will set off the fireworks whenever they hear the fireworks from the northside. At 9:00 on the northside, both the planned fireworks display and another, unplanned, fireworks display go off simultaneously. Either one on its own would be noisy enough to get the southside fireworks display going. We can model this case with $\overline{S} := \overline{P} \wedge \neg U$ (that is: the *s*outhside fireworks will *not* go off ($\overline{S} = 1$) iff both the *p*lanned fireworks *don't* go off ($\overline{P} = 1$) and the *u*nplanned fireworks don't go off ($\neg U = 1$). If we assume that a planned fireworks display is default and an unplanned one deviant, then we have an isomorphism between the two models. But the planned fireworks on the northside was (at least a part of) a cause of the planned fireworks on the southside, and Watson's antidote was not a cause of Holmes's survival. Cf. Hiddleston (2005).

This is not an easy choice. At the moment, I am most inclined to take the third option. As I see it, the only way of making the first option palatable is to provide an error theory for the mistaken causal judgement. That is: it would require explaining away either our judgement that Fozzie Bear didn't cause the money to be incinerated in *Self-Underminer*, or our judgement that Crazy Harry caused the money to be incinerated in *Preempter*. Lewis (2004) offers an error theory like this. His theory says that Fozzie Bear *did* cause the money to be incinerated in *Self-Underminer*, and in his defense, he suggests that we only incline to the contrary judgement because "[w]e note that [Fozzie's action] didn't matter; [the money would have been incinerated] all the more easily without it. The effect doesn't depend on the cause. The idea that causation requires whether-whether dependence may retain some grip on us." I have a hard time accepting this as an explanation for our differing judgements, for it points to a feature which *Preempter* and *Self-Underminer* have in common. The money's incineration doesn't depend upon either Fozzie's or Harry's act. If the lack of dependence is enough to confound us and corrupt our causal judgements, why isn't our judgement about *Preempter* similarly corrupted?

Another very natural thought is that the reason we are not inclined to say that Fozzie caused the incineration is that Fozzie didn't *intend* to incinerate the money, and therefore isn't *morally* responsible for the incineration. That's an important asymmetry between Fozzie and Harry, and it's easy to understand how this could influence our causal judgements. For it is plausible that, in general, we have a tendency to conflate causal and moral responsibility. My own view is that you are morally responsible for the foreseeable effects of your wrong choices.[9] On the assumption that what Fozzie did wasn't wrong, he's not morally responsible, whether or not he's causally responsible. I think this error theory is *prima facie* plausible, but unfortunately, I don't think it stands up to scrutiny. The reason is that altering whether Fozzie has made a morally wrong choice doesn't seem to blunt the intuition that he didn't cause the incineration. Suppose, for instance, that both Fozzie and Harry want to watch the money burn, but Fozzie is misinformed, and thinks that this will only happen if they both flip their switches. So Fozzie flips his switch with the intention of incinerating the money. Harry follows suit, and the money is incinerated. In this version of the case, I'm inclined to say that, given Fozzie's beliefs, his decision to flip the switch was morally wrong (think about how many people could be helped with that money), but nonetheless that, while Fozzie *thinks* that his action was a cause of the money's incineration, he is wrong about that. He didn't accomplish anything, in spite of the fact that his decision to flip the switch was morally wrong. While he's morally responsible and blameworthy for *attempting* to incinerate the money, he's not morally responsible for the money's incineration, because in spite of his attempts, he didn't cause the incineration.

The second option for avoiding the contradiction raises pressing methodological questions which I do not see how to answer. If a theory of causation is not invariant under the removal of interpolated variables, then how will we ever know whether we've included *enough* interpolated variables in our model of a system? For almost every causal model, and almost any pair of parent and child variables in the model, $Pa \rightarrow Ch$,

9. See Moore (2009) and Sartorio (2016) for more on the relationship between causal and moral responsibility.

we will be able to enrich our model by interpolating an additional variable in between *Pa* and *Ch*. If doing so may change the verdict of our theory, then how many variables must we interpolate along any given path in the model, before we can trust the verdicts of our theory? This is not a rhetorical question, but a genuine one. There may be satisfactory answers, but I haven't found any that satisfy me, and this inclines me to reject the second option. (But I would be excited to see others explore it further.)

So I am left with the third option. A causal model does not tell us enough to determine whether one variable's value is a cause of another's—not even when we include information about which variable values represent default states and which represent deviations therefrom. There is something missing—something else which we must know in order to determine whether *c* is a cause of *e*.

3 | NEEDS

As I foreshadowed in the introduction, I will suggest that the missing piece of the puzzle is whether *e already had all it needed* to happen without *c*. On the view I will develop here, *c* is a cause of *e* only if *c* is a *deviant* occurrence, and *c* meets a *need* of *e*'s. So, on this view, default states can never be causes, and a deviant event, *c*, can only be a cause of *e* when *c* meets one of *e*'s needs. The reason why Fozzie Bear didn't cause the incineration in *Self-Underminer* is that, even though Fozzie's flipping the switch was a deviation from the default state of doing nothing, the incineration didn't *need* Fozzie's flip in order to happen. It had all it needed to happen either way. And the reason why Crazy Harry did cause the incineration in *Preempter* is that Harry's flipping the switch was a deviation from the default state of doing nothing, and, moreover, Harry's flipping the switch met one of the incineration's needs. (Of course, had Harry not met this need, Waldorf would have stepped in and given the incineration all it needed to occur. Even so, it was Harry, and not Waldorf, who met the need.)

One natural way of thinking about what an effect *needs* in order to occur is in terms of dependence: *c* met a need of *e*'s iff, had *c* not happened, *e* wouldn't have happened, either. But this understanding would do nothing to distinguish *Self-Underminer* from *Preempter*. Instead, I will suggest an alternative understanding which focuses first-and-foremost, not on dependence, but rather on the propagation of deviancy through a system. I will suggest that the reason why Fozzie didn't meet any of the incineration's needs is that the incineration had already received all the deviancy it needed to occur without Fozzie's action. And the reason why Harry did meet one of the incineration's needs is that, without Harry's action, the incineration would have required *additional* deviancy (from Waldorf) in order to occur.

That's a very rough-and-ready characterisation. I'll spend the rest of this section getting more precise about it. I'll begin by introducing the notion of a *process*. Consider the following case:

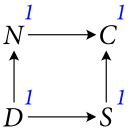
> *Collision*
> The north and south tracks each have switches directing trains onto the central track. Gallant is stationed at the switch for the north track, and Goofus is stationed at the switch for the south. The dispatcher sends the order: "direct the north train onto the central track". Gallant flips his switch, as ordered. Unfortunately, Goofus has forgotten which track he's

> stationed at, and he flips his switch, too. The north and south trains collide.

The consequences of the dispatcher's order propagate out through the world. A given route along which they propagate is what I am calling a *process*. As I use the term, there are three processes leading from the dispatcher's orders to the collision. One propagates from the dispatcher's order to Gallant's flipping the north switch, on to the north train being directed to the central track, and finally to the collision. Another propagates from the dispatcher's order to the collision *via* Goofus's flipping the south switch and the south train being directed onto the central track. And a *third* propagates from the dispatcher's orders to the collision *via* both Gallant's and Goofus's actions.

We can characterise this notion of a process with the aid of structural equations modelling. For instance, we may model *Collision* with this system of equations:

$$\begin{aligned} D &= 1 \\ N &:= D \\ S &:= D \\ C &:= N \wedge S \end{aligned}$$



Here, $D = 1$ represents the dispatcher giving the order, and $D = 0$ stands for the dispatcher doing nothing. $N = 1$ stands for Gallant flipping the north switch, and $N = 0$ stands for him doing nothing. $S = 1$ stands for Goofus flipping the south switch, and $S = 0$ stands for him doing nothing. Finally, $C = 1$ stands for the north and south trains colliding, and $C = 0$ stands for them not colliding. The equations tell us that the collision will only take place if both Gallant and Goofus flip their switches, and that each of Gallant and Goofus will flip their switch iff the dispatcher gives the order.

To the right of the system of equations, I've given the associated directed graph, which consists of four directed edges: $D \to N, D \to S, N \to C$, and $S \to C$. In a graph like this, a *directed path* from one variable, X, to another variable, Y, is a collection of directed edges which lead from X to Y when they are aligned tail-to-tip.[10] For instance, in the model of *Collision*, both $\{D \to N, N \to C\}$ and $\{D \to S, S \to C\}$ are directed paths from D to C. For ease of presentation, I'll abbreviate the first of these directed paths with '$D \to N \to C$' and the second with '$D \to S \to C$'. What I will call a *network* from X to Y is just the union of any number of directed paths from X to Y. For instance, the union of the directed paths $D \to N \to C$ and $D \to S \to C$ is a network from D to C. For ease of presentation, I'll write this network '$D \to N \to C \leftarrow S \leftarrow D$'. Take some network, $\mathcal{N}$, and suppose that the directed edge $U \to V \in \mathcal{N}$. Then, I will say that U is one of V's *parents* along the network $\mathcal{N}$—or, alternatively, U is one of V's $\mathcal{N}$-parents.

Take any variable, V, along a network, $\mathcal{N}$, leading from X to Y, such that $V \neq X$. Let '**P**' be V's $\mathcal{N}$-parents. And let '**O**' be the *o*ther variables showing up on the right-hand-side of V's structural equation. Then, V's structural equation has the form

10. More carefully, a set of directed edges is a *directed path* iff there's a sequence of variables $(X_1, X_2, \ldots, X_N)$ such that (a) no two variables in the sequence are identical; (b) for each $i \in \{1, 2, \ldots, N-1\}$, $X_i \to X_{i+1}$ is included in the set; and (c) nothing else is included.

$V := f(\mathbf{P}, \mathbf{O})$, for some function f of the variables in $\mathbf{P}$ and $\mathbf{O}$. Let v be the actual value of V, and let $\mathbf{p}$ and $\mathbf{o}$ be the actual assignments of values to the variables in $\mathbf{P}$ and $\mathbf{O}$, respectively.[11] Then, it must be that $f(\mathbf{p}, \mathbf{o}) = v$. Now, I'll say that $V = v$, rather than v^*, *locally depends upon* $\mathbf{P} = \mathbf{p}$, rather than $\mathbf{p}^*$, iff $f(\mathbf{p}^*, \mathbf{o}) = v^*$. That is, V *locally depends* upon $\mathbf{P}$ iff, when we confine attention to V's structural equation, wiggling the values of $\mathbf{P}$ wiggles the value of V. Local dependence is not quite the same as counterfactual dependence. To appreciate the difference, consider the model of *Thwarted Assassination* from §2. In that model, $D = 0$, rather than 1, does not counterfactually depend upon $M = 1$, rather than 0—that is: had Moriarty not given the encoded order to poison Holmes, Holmes would still have lived. However, $D = 0$, rather than 1, does *locally* depend upon $M = 1$, rather than 0—that is, Holmes's survival does *locally* depend upon Moriarty's coded message; given that the drink was poisoned, had Moriarty's order not alerted Holmes, he would have died. In many cases, local dependence and counterfactual dependence come to the same thing. When they do, I'll just talk about 'dependence', without bothering to distinguish between the two.

With these notions of a network and local dependence in hand, we may provide a characterisation of a *process* in terms of a correct system of structural equations. To see whether there is a process leading from a variable C to another variable E, take any correct system of structural equations which includes the variables C and E.[12] Then, suppose that there is a network leading from C to E with the following property: there is a way of assigning each variable lying along the network a *contrast* value such that (a) E's contrast is different from its actual value, and (b) for each variable $D \neq C$ lying along the network, with actual value d and assigned contrast d^*, $D = d$, rather than $D = d^*$, locally depends upon its $\mathcal{N}$-parents' values, rather than their contrasts. If that's so, then we have a *dependence network* leading from $C = c$, rather than c^*, to $E = e$, rather than e^*, *via* the network $\mathcal{N}$. Any dependence network leading from C to E represents a *process* by which C's value, rather than its contrast, is propagated to E's value, rather than its contrast.

Dependence Network Within a structural equations model, a *dependence network* leading from $C = c$, rather than c^*, to $E = e$, rather than e^*, is a network leading from C to E, $\mathcal{N}$, and an assignment of contrasts to the variables appearing along $\mathcal{N}$ such that:

(a) C's value is c, its contrast is c^*, and $c \neq c^*$;

(b) E's value is e, its contrast e^*, and $e \neq e^*$; and

(c) for every $D \neq C$ along $\mathcal{N}$, D's value, rather than its contrast, locally depends upon its $\mathcal{N}$-parents' values, rather than their contrasts.

After dotting all the 'i's and crossing all the 't's, that's a bit of a mouthful, but I hope that the underlying idea is intuitive enough. A dependence network from C to E is just a (perhaps branching) chain of local dependence leading from C to E. For instance,

11. An *assignment* of values to $\mathbf{P}$, $\mathbf{p}$, is just a function from the variables in $\mathbf{P}$ to real numbers. You hand the assignment $\mathbf{p}$ a variable $P \in \mathbf{P}$, and it hands you back one of P's values. I use '$\mathbf{P} = \mathbf{p}$' as an abbreviation for '$(\forall P \in \mathbf{P})\ P = \mathbf{p}(P)$'.

12. See GALLOW (ms) for more on what makes a system of structural equations *correct*.

consider again the case of *Collision*. In the system of structural equations provided above, $D \rightarrow N \rightarrow C$ gives us a dependence network when each variable is assigned the contrast 0. For $C = 1$, rather than $C = 0$, depends upon $N = 1$, rather than $N = 0$. And $N = 1$, rather than $N = 0$, depends upon $D = 1$, rather than $D = 0$. For this reason, I say that there is a *process* leading from the dispatcher's order to the collision, *via* Gallant's flipping the switch on the north track. Similarly, $D \rightarrow N \rightarrow C \leftarrow S \leftarrow D$ gives us a dependence network when each variable is assigned the contrast value 0. For $C = 1$, rather than $C = 0$, depends upon $N = 1 \wedge S = 1$, rather than $N = 0 \wedge S = 0$. Likewise, $N = 1$, rather than $N = 0$, depends upon $D = 1$, rather than $D = 0$; and $S = 1$, rather than $S = 0$, depends upon $D = 1$, rather than $D = 0$. So there is a process leading from the dispatcher's order to the collision *via* both Gallant's and Goofus's action.

Process There is a *process* leading from c, rather than c^*, to e, rather than e^*, iff, according to a correct structural equations model, there is a dependence network leading from $C = c$, rather than c^*, to $E = e$, rather than e^*.

I've introduced these *two* terms ('dependence network' and 'process'), to mark the distinction between representation and reality. A system of equations *represents* some chunk of the world's causal structure. A dependence network is a part of this representation. A process, on the other hand, is the thing in reality to which that part of the representation corresponds. It's important to notice that both dependence networks and processes are associated with a series of *contrasts*. In the representation, these contrasts are alternative values of variables. In reality, the contrasts are alternative ways the parts of the world represented by variables could be. These contrasts are an integral part of the process; different contrasts make for different processes. For illustration, suppose that, as the campfire burns, you throw potassium chloride into the flames.[13] Potassium chloride turns the flames purple. You also have copper sulfate, which would have turned the flames green. Then, there are at least *two* processes connecting your action to the event of the flames turning purple. For we can either assign your action the contrast of doing nothing (the default), or we can assign it the contrast of adding copper sulfate. Likewise, we can assign the flames' turning purple the contrast of them remaining yellow (the default), or the contrast of them turning green. In either case, we will have a chain of dependence leading from your act to the flames turning purple. For, had you done nothing instead of adding potassium chloride, the flames would have remained yellow, instead of turning purple. And, had you added copper sulfate instead of potassium chloride, the flames would have turned green instead of purple. These different contrasts make for different processes. These two processes are importantly different, since, along the first of these processes, your action transmits deviancy to the flames' color. That is: it is *via* the first process that the deviancy of your adding potassium chloride (rather than not) is transmitted to the flames' color. The deviancy of their being purple (rather than not) depends upon the deviancy of your adding potassium chloride (rather than not). On the other hand, the second process does not transmit deviancy to the flames' color. For—I am supposing—adding potassium chlo-

13. I borrow this kind of example from Woodward (1984).

ride is no more deviant than adding copper sulfate, and the flames' turning purple is no more deviant than their turning green.

On the view I am developing here, causation is closely tied to the propagation of deviant behaviour. What it is for c to cause e is for c to be deviant, and for this deviancy to propagate from c to e. And, in order for c's deviancy to *propagate* to e, c and e must be connected by a process, $\mathcal{P}$, on which (i) c is deviant, and no more default than its contrast, so that c has deviancy to transmit through $\mathcal{P}$; and (ii) the deviancy c provides through $\mathcal{P}$ is *needed* for e to happen—or, putting the same point in different terms: without the deviancy c transmits through $\mathcal{P}$, the remaining deviancy was not enough for e to happen.

Suppose you have a process, $\mathcal{P}$, running from c to e. Associated with this process will be some contrasts for c and e—call them 'c^*' and 'e^*'. On my view, in order for c to be eligible as a cause of e, c must be deviant, and c^* must be no more deviant than c, so let's assume that this is so. Now, in order to see whether the deviancy c transmitted through $\mathcal{P}$ was *needed* for e to happen, we must see whether, without that deviancy, the remaining deviancy would have been enough for e to happen or not. So we must consider what things would have been like, outside of the process $\mathcal{P}$, had c^* happened instead of c. Some things outside of the process $\mathcal{P}$ will be completely unaffected by whether c or c^* happens. For our purposes, they can be safely ignored. Other things outside of $\mathcal{P}$ may have been *less* deviant, had c^* happened. If so, this shows us that c has transmitted deviancy along additional processes besides $\mathcal{P}$. We should not hold fixed the transmission of this deviancy. This is not deviancy that e would have received with or without c—it is deviancy directly attributable to c. Without c, e would not have had this deviancy. Finally, there may be things outside of $\mathcal{P}$ which would have been *no less* deviant, had c^* happened. These are c's *non-deviant consequences, external to* $\mathcal{P}$. They are the consequences of c's happening in place of c^* which did not involve the transmission of additional deviancy.

Non-Deviant External Consequences Given a process beginning with c, rather than c^*, n is a *non-deviant external consequence* iff (a) n is not a part of the process, (b) were c^* to happen, n^* would have happened instead of n, and (c) n is no more deviant than n^*.

(Non-deviant external consequences sub-divide into two kinds: (i) consequences which would have been *more* deviant, had c^* happened, and (ii) those which would have been altered, but *just as* deviant, had c^* happened. Things in the first category were *inhibited* by c's happening in place of c^*. They represent respects in which c has potentially *deprived* e of deviancy being transmitted through other processes. Things in the second category were *altered* by c happening instead of c^*, but this alteration does not supply any additional deviancy to e.)

To determine whether e had all the deviancy it needed without the deviancy c transmitted through $\mathcal{P}$, ask yourself this: would e have happened, were c^* to happen while all of the non-deviant external consequences of c were held fixed? If so, then the deviancy c provided through $\mathcal{P}$ was not needed for e to happen. If not, then it was.

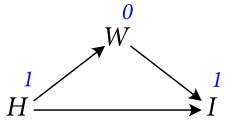
Needs Suppose there is a process, $\mathcal{P}$, running from c, rather than c^*, to e. And suppose that c is deviant, and c^* is no more deviant than c. Then, e *needs* the deviancy transmitted from c through $\mathcal{P}$ iff: were c^* to happen, and were every non-deviant consequence of c external to $\mathcal{P}$ to be held fixed, e would not happen.

4 | CAUSATION

On the theory I am proposing, c is a cause of e only if there is a process, $\mathcal{P}$, connecting c to e such that (a) c is deviant and no more default than its contrast, so that c transmits deviancy through $\mathcal{P}$; and (b) e *needs* the deviancy that c transmits through $\mathcal{P}$. There's an additional condition needed to transform this 'only if' into an 'if and only if', but it won't be relevant to any of the cases we will consider here, so I will ignore it. In all of the cases we'll consider here, conditions (a) and (b) will be both necessary and sufficient for c being a cause of e. (I provide the full theory in the appendix.)

Let's apply this theory to our examples. Start with the case of *Preempter*. The causal model of the case, $\mathcal{M}_P$, is reproduced below.

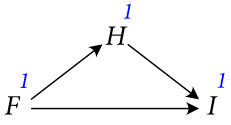
$$\begin{aligned} H &= 1 \\ W &:= \neg H \\ I &:= [H \neq W] \end{aligned}$$



In this model, $H \rightarrow I$ is a dependence network leading from $H = 1$, rather than 0, to $I = 1$, rather than 0. For, when we look just at I's structural equation, with the actual values of H and W, $I := [1_H \neq 0_W]$,[14] wiggling H from 1 to 0 wiggles I from 1 to 0. So $I = 1$, rather than 0, locally depends upon $H = 1$, rather than 0. This tells us that there is a process leading from Harry's flipping the switch to the money's being incinerated. Moreover, Harry's flipping the switch ($H = 1$) is more deviant than his doing nothing ($H = 0$). So Harry's action transmits deviancy through this process. And this is deviancy which the money's incineration *needed*. For Waldorf's doing nothing is a non-deviant external consequence of Harry's flipping the switch. And, had Harry not flipped the switch while Waldorf continued to do nothing, the money would not have been incinerated. So Harry's flipping the switch is a cause of the money's incineration.

Contrast this with the case of *Self-Underminer*. The causal model of *Self-Underminer*, $\mathcal{M}_S$, is reproduced below.

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



In this model, there are two dependence networks leading from $F = 1$, rather than 0, to $I = 1$, rather than 0. Firstly, there's the dependence network leading from F directly to I, which bypasses the variable H. For when we look just at I's structural equation, with the actual values of F and H, $I := [1_F = 1_H]$, wiggling F from 1 to 0 wiggles I from 1 to 0. So there is a direct process transmitting the deviancy of Fozzie's flipping the switch to the money's incineration. There are no non-deviant consequences of Fozzie's flipping the switch which are external to this process, so to see whether the incineration needed the deviancy which Fozzie's flip provided along this process, we must ask whether the incineration would have happened, had Fozzie not flipped. And

14. I'm subscripting '1' and '0' with 'H' and 'W', respectively, just to remind the reader of which variables they are the values *of*.

the answer is 'yes, it would still have happened'. So Fozzie's flip was not needed, and we may conclude that Fozzie did not cause the money's incineration *via* the process which bypasses Crazy Harry's flip.

But there's another dependence network in the model $\mathcal{M}_S$: the one which goes *via* the path $F \rightarrow H \rightarrow I$. For $I = 1$, rather than 0, depends upon $H = 1$, rather than 0. And $H = 1$, rather than 0, depends upon $F = 1$, rather than 0. But, in this case, too, there are no non-deviant external consequences of Fozzie's flipping, rather than not. And, holding all of the none of them fixed, had Fozzie not flipped, the money would still have been incinerated. So Fozzie's flip was not needed—the money's incineration already had everything it needed to happen without his action. And so Fozzie's flip is not a cause of the incineration *via* the process which goes by way of Harry's flip. Since there are no other processes leading from Fozzie's flip to the explosion, Fozzie didn't cause it *via* any process at all.

It's natural to look at this treatment of *Self-Underminer* and think that it depends upon an irrelevant feature of the case. For it's natural to think that you could easily insert into the story some non-deviant external consequence of Fozzie's flip on which the incineration depends, and that this wouldn't make any difference to whether Fozzie caused the money to be incinerated. I agree that we can easily add non-deviant external consequences to the story, but I will contend that—even though it may not be obvious at first—those non-deviant external consequences *do* make a causal difference. Once the story has been so emended, there's no longer anything distinguishing the case from *Preempter*.

To illustrate the objection, consider the following variant of *Self-Underminer*, which I will call—perhaps presumptuously—*Preempter**: Unbeknownst to Fozzie and Harry, Waldorf and Statler have each installed bombs beneath the money. Fozzie Bear's switch controls Waldorf's bomb, and Crazy Harry's switch controls Statler's bomb. If either of these switches is set to Left, then the bomb it controls will be armed. If the switch is set to Right, then the bomb it controls will be *dis*armed. If both bombs are armed, then there will be a power surge, and neither bomb will explode. With this set-up, suppose that everything proceeds exactly as before. Fozzie flips his switch to Right, disarming Waldorf's bomb. In response, Crazy Harry flips his switch to Left, arming Statler's bomb. Statler's bomb explodes, and the money is incinerated. When it comes to the patterns of counterfactual dependence between the two switches and the incineration, there is no difference between this case and *Self-Underminer*. The money will be incinerated iff the switches are in opposite positions, and the money will be saved iff the switches are in the same positions.

Admittedly, the intuitions here are slippery. *Preempter** is a bit of a duck-rabbit. When I go from thinking about *Self-Underminer* to thinking about *Preempter**, I'm most inclined to say that Fozzie didn't cause the money to be incinerated; and when I go from thinking about *Preempter* to thinking about *Preempter**, I'm most inclined to say that he did. This new case sits uneasily between the first two, and my intuition is uncertain how to deal with it. (Incidentally, notice that *Preempter** is modelled by $\mathcal{M}_X$, from §2, if we let X stand for whether Fozzie flips, W for whether Waldorf's bomb is armed, and H for whether Statler's bomb is armed.) I would not be surprised if the reader saw this as further reason to think that we should not have distinguished between *Self-Underminer* and *Preempter* in the first place. However, I continue to be moved by the arguments I presented against this reaction back in §2.

My considered view is that, to the extent that we are still inclined to say that Fozzie Bear did not cause the money to be incinerated in *Preempter**, this inclination has two sources. Firstly: recognising the similarity between this case and *Self-Underminer*, we may be overly inclined to ignore the details about how the switches' positions bring about the incineration. That is, when thinking about the case, we may be overly inclined to 'black box' away these details as irrelevant, focusing just on the patterns of counterfactual dependence between the switches and the explosion. It's well understood that 'black boxing' away details in this way can make a difference to our causal judgements. For an illustration of this phenomenon, suppose that Hugh took two sleeping pills—one pink, one purple—and fell asleep. Both of the pills are soporific, and either on its own would have brought on sleep. Taking both has no negative consequences, and Hugh would get just as sleepy after taking both as he would if he'd only taken one. Given this information, most of us are not inclined to draw any causal distinction between the pills. We're inclined to say that the pink pill was a cause of Hugh's sleepiness iff the purple pill was, too. But suppose I go on to tell you that the pink pill contains a chemical which breaks down the active ingredient in the purple pill—so that, after the pink and purple pills had dissolved in Hugh's stomach, only the active ingredient from the pink pill remained intact. When we attend closely to this additional information, most of us are now inclined to insist that, while the pink pill was a cause of Hugh's sleepiness, the purple pill was not.[15] (In my view, cases like this teach us that our causal judgements about the stories described in vignettes can be overly hasty. In this respect, our causal judgements are not unlike our moral judgements. Nonetheless, under the assumption that the vignette tells us everything relevant there is to tell, I continue to think that our considered causal judgements tend to track the causal facts.)

Even when we are careful to direct our attention to Waldorf and Statler's bombs and their relationship to the two switches, there remains some inclination to deny that Fozzie caused the incineration (though I expect the strength of this inclination to vary from reader to reader). I suspect this lingering inclination derives its force from a conflation between *moral* and *causal* responsibility. Back in §2, I considered an error theory of our judgement in *Self-Underminer* which appealed to such a conflation. I rejected this error theory because the judgements in that case did not appear sensitive to changes in Fozzie's intentions. When it comes to *Preempter**, however, matters are different. In this case, the degree to which we are inclined to count Fozzie Bear's action as a cause of the incineration *is* sensitive to changes in Fozzie's intentions. To illustrate this point, suppose that both Fozzie and Harry want the money to burn, both know about the existence of Statler's bomb, and both know that it will explode iff Fozzie's flip is set to Right and Harry's set to Left. However, they falsely believes that Statler's bomb is the only bomb. (Thus, if Fozzie hadn't flipped his switch, Harry wouldn't have bothered to flip his, either.) So Fozzie flips his switch. This has the unintended consequence of deactivating Waldorf's bomb, and the intended consequences of Harry flipping his switch, and Statler's bomb being armed. And Statler's bomb, once armed, incinerates the money. With these changes made, I no longer see any important difference between this case and *Preempter*. Waldorf's arming of his bomb was a backup,

15. See Paul & Hall (2013, §3.6) for more on 'black box cases'.

would-be cause of the incineration, but this backup was preempted by Fozzie's flipping his switch to Right. At the same time that Fozzie's flip deactivated Waldorf's bomb, it also initiated a sequence of events which culminated in Statler's bomb incinerating the money. Waldorf's bomb is a backup, would-be cause of the money's incineration, preempted by Fozzie Bear's flip.

At least, that's how it seems to me. If you're still on the fence, the I can offer you an argument that Fozzie Bear is causally responsible in *Preempter**. This argument relies upon the view that people are only morally responsible for the effects of their choices. In the variant of the case in which Fozzie intends to incinerate the money, it appears that Fozzie is morally responsible for the incineration. But if he did not cause the money to be incinerated, then the incineration cannot be an *effect* of his choice to flip the switch, and so he cannot be held morally responsible.

As I'll use the term here, a *causal model*, $\mathcal{M}$, is a 5-tuple, $(\mathcal{S}, \mathcal{E}, \mathbf{u}, \succeq, \mathcal{I})$. $\mathcal{S}$, $\mathcal{E}$, and $\mathbf{u}$, are familiar fare from the causal modelling literature. $\mathcal{S}$ is a *signature—i.e.*, a triple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ a set of endogenous variables, and $\mathcal{R}$ is a function from each $X \in \mathcal{U} \cup \mathcal{V}$ to the *range* of X, $\mathcal{R}(X)$, which is the set of possible values X can take on. Given a *set* of variables $\mathbf{X}$, we may define $\mathcal{R}(\mathbf{X})$ to be the set of possible assignments of values to the variables in $\mathbf{X}$. $\mathcal{E}$ is a system of *structural equations*, with one structural equation for each endogenous variable $V \in \mathcal{V}$, and $\mathbf{u}$ is an assignment of values to the variables in $\mathcal{U}$. $\succeq$ is a *deviancy ordering*. Formally, it is a function from some tuple of variables $\mathbf{X}$ to a pre-order, $\succeq_{\mathbf{X}}$, over the values in $\mathcal{R}(\mathbf{X})$. The interpretation of this pre-order is that $\mathbf{x} \succeq_{\mathbf{X}} \mathbf{x}^*$ iff $\mathbf{X} = \mathbf{x}$ is no less deviant than $\mathbf{X} = \mathbf{x}^*$. The least elements of $\succeq_{\mathbf{X}}$ are the *default* values of $\mathbf{X}$, and all other assignments are deviations therefrom.[16] I will return to $\mathcal{I}$ momentarily.

I will allow *sets* of variable values to serve as causes.[17] So I must generalise the notion of a *network* which I provided in the main text above. Say that a network from a *set* of variables $\mathbf{C}$ to E is a union of directed paths from some $C \in \mathbf{C}$ to E. The definition of a *dependence network* remains the same, once 'C', 'c', and 'c^*' are exchanged for '$\mathbf{C}$', '$\mathbf{c}$', and '$\mathbf{c}^*$'. Notice that, once a network $\mathcal{N}$ is specified, all we need in order to work out the contrast value for every variable along the network are the contrasts for each $C \in \mathbf{C}$. For condition (c) in the definition of a dependence network assures us that, for any *other* variable on the network $V \notin \mathbf{C}$, its contrast is the value v^* such that $f(\mathbf{p}^*, \mathbf{o}) = v^*$ (where f is the function on the right-hand-side of V's structural equation, $\mathbf{P}$ are V's $\mathcal{N}$-parents, $\mathbf{p}^*$ are their designated contrasts, and $\mathbf{o}$ are the actual values of V's *o*ther parents). So, once the contrasts for $\mathbf{C}$ are settled, every other contrast in the network is settled as well. So we can refer to a dependence network with a pair $(\mathcal{N}, \mathbf{c}^*)$, where $\mathbf{c}^*$ are the contrasts assigned to the variables at the start of the network $\mathcal{N}$.

Suppose that two variables, D and R, lie along a network $\mathcal{N}$, and that there is a directed path *o*utside of the network, $\mathcal{O} : D \rightarrow O_1 \rightarrow O_2 \rightarrow \cdots \rightarrow O_N \rightarrow R$, such that none of the directed edges in $\mathcal{O}$ appear in $\mathcal{N}$. Then, call D a *departure* variable, and call 'R' a *return* variable (relative to $\mathcal{N}$).

Now, we can return to the final component of a causal model: $\mathcal{I}$. As I explained in §3 above, a dependence network from $\mathbf{C}$ to E corresponds to a *process* from $\mathbf{C}$ to E. And, given any process from $\mathbf{C}$ to E, there will be the value which E would have had, had $\mathbf{C}$ been $\mathbf{c}^*$ and every non-deviant external consequence of $\mathbf{C} = \mathbf{c}$, rather than $\mathbf{c}^*$, been held fixed. Let us call that value E's *inertial* value. $\mathbf{C} = \mathbf{c}$ will meet a *need* of $E = e$ along the process represented by $(\mathcal{N}, \mathbf{c}^*)$ iff e is not E's inertial value, relative to that process. Notice that there is no need to confine our attention to the variables which lie at the *end* of the process. We could use the same definition to determine which values are inertial for any *other* variables along the process. Take some variable, V, which lies along the process corresponding to $(\mathcal{N}, \mathbf{c}^*)$, and ask yourself: had $\mathbf{C} = \mathbf{c}^*$ happened and had every non-deviant consequence of $\mathbf{C} = \mathbf{c}$ external to the process been held fixed, which value would V have taken on? Whichever value that is, it is the *inertial*

16. As usual, we define $\mathbf{x} \succ_{\mathbf{X}} \mathbf{x}^* \stackrel{\text{def}}{=} \mathbf{x} \succeq_{\mathbf{X}} \mathbf{x}^* \wedge \neg \mathbf{x}^* \succeq_{\mathbf{X}} \mathbf{x}$.

17. See GALLOW (forthcoming, §3).

value of v, relative to that process.

The point of the function $\mathcal{I}$ is to tell us which values of variables lying along a process are *inertial*, and which are not. Formally, it is a function from a dependence network, $(\mathcal{N}, \mathbf{c}^*)$, and a variable V along that network, to a value, $v' \in \mathcal{R}(V)$. The interpretation is that $\mathcal{I}(\mathcal{N}, \mathbf{c}^*, V) = v'$ iff $v^* = \prime$ is the *inertial* value of V, relative to the process corresponding to the dependence network $(\mathcal{N}, \mathbf{c}^*)$.[18]

With this all in place, we may carefully state the theory:

Causation In a causal model $\mathcal{M} = (\mathcal{S}, \mathcal{E}, \mathbf{u}, \geq, \mathcal{I})$, $\mathbf{C} = \mathbf{c}$, rather than $\mathbf{c}^*$, is a cause of $E = e$, rather than e^*, iff there is some dependence network, $(\mathcal{N}, \mathbf{c}^*)$, leading from $\mathbf{C} = \mathbf{c}$, rather than $\mathbf{c}^*$, to $E = e$, rather than e^*, such that:

(a) $\mathbf{C} = \mathbf{c}$ is deviant, and $\mathbf{C} = \mathbf{c}^*$ is no more deviant than $\mathbf{C} = \mathbf{c}$;

(b) $E = e$ needed the deviancy $\mathbf{C} = \mathbf{c}$ provided *via* $(\mathcal{N}, \mathbf{c}^*)$, $\mathcal{I}(\mathcal{N}, \mathbf{c}^*, E) \neq e$;

(c) for any departure and return variables along $\mathcal{N}$, D and R:

(c1) D's value is more deviant than its contrast;

(c2) R's value needed the deviancy $\mathbf{C}$ provided *via* $(\mathcal{N}, \mathbf{c}^*)$, $\mathcal{I}(\mathcal{N}, \mathbf{c}^*, R) \neq r$; and

(d) $(\mathcal{N}, \mathbf{c}^*)$ is *minimal*; that is, there is no proper sub-network $\mathcal{N}' \subset \mathcal{N}$, beginning with a subset $\mathbf{C}' \subseteq \mathbf{C}$, such that $(\mathcal{N}', \mathbf{c}^*|_{\mathbf{C}'})$ is a dependence network satisfying conditions (a), (b), and (c) above.[19]

REFERENCES

Collins, J., N. Hall & L. A. Paul, editors. 2004. *Causation and Counterfactuals*. The MIT Press, Cambridge, MA. [iii]

Gallow, J. Dmitri. 2016. "A Theory of Structural Determination." *Philosophical Studies*, vol. 173 (1): 159–186. [3]

—. forthcoming. "A Model-Invariant Theory of Causation." *Philosophical Review*. [2], [7], [i]

—. ms. "Model-Variance in Theories of Token Causation." [3], [7], [13]

Hall, Ned. 2007. "Structural Equations and Causation." *Philosophical Studies*, vol. 132 (1): 109–136. [3], [4], [5], [6]

Halpern, Joseph Y. 2008. "Defaults and Normality in Causal Structures." *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning*, 198–208. [5]

—. 2016. *Actual Causality*. MIT Press, Cambridge, MA. [5]

18. If $(\mathcal{N}, \mathbf{c}^*)$ is not a dependence network or V does not lie along $\mathcal{N}$, then $\mathcal{I}(\mathcal{N}, \mathbf{c}^*, V)$ is undefined.

19. '$\mathbf{c}^*|_{\mathbf{C}'}$' is the assignment $\mathbf{c}^*$, restricted to the variables in $\mathbf{C}'$.

Halpern, Joseph Y. & Christopher Hitchcock. 2015. "Graded Causation and Defaults." *The British Journal for the Philosophy of Science*, vol. 66 (2): 413–457. [5]

Halpern, Joseph Y. & Judea Pearl. 2005. "Causes and Explanations: A Structural-Model Approach. Part 1: Causes." *The British Journal for the Philosophy of Science*, vol. 56: 843–887. [3]

Hiddleston, Eric. 2005. "Causal Powers." *The British Journal for the Philosophy of Science*, vol. 56: 27–59. [4], [5], [9]

Hitchcock, Christopher. 1996a. "Farewell to Binary Causation." *Canadian Journal of Philosophy*, vol. 26 (2): 267–282. [4]

—. 1996b. "The Role of Contrast in Causal and Explanatory Claims." *Synthese*, vol. 107 (3): 395–419. [4]

—. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs." *The Journal of Philosophy*, vol. 98 (6): 273–299. [3]

—. 2007. "Prevention, Preemption, and the Principle of Sufficient Reason." *Philosophical Review*, vol. 116 (4): 495–532. [3], [5]

—. 2011. "Trumping and contrastive causation." *Synthese*, vol. 181: 227–240. [4]

Hitchcock, Christopher & Joshua Knobe. 2009. "Cause and Norm." *Journal of Philosophy*, vol. 106 (11): 587–612. [5], [9]

Kahneman, Daniel & Dale T. Miller. 1986. "Norm Theory: Comparing Reality to Its Alternatives." *Psychological Review*, vol. 94 (2): 136–153. [5]

Lewis, David K. 1973. "Causation." *The Journal of Philosophy*, vol. 70 (17): 556–567. [4]

—. 2004. "Causation as Influence." In Collins et al. (2004), chap. 3, 75–106. [1], [2], [10]

Maslen, Cei. 2004. "Causes, contrasts, and the nontransitivity of causation." In Collins et al. (2004), 341–357. [4]

McDermott, Michael. 1995. "Redundant Causation." *The British Journal for the Philosophy of Science*, vol. 46 (4): 523–544. [1]

McGrath, Sarah. 2005. "Causation by Omission: A Dilemma." *Philosophical Studies*, vol. 123: 125–148. [5], [9]

Moore, Michael. 2009. *Causation and Responsibility*. Oxford University Press, Oxford. [10]

Paul, L. A. & Ned Hall. 2013. *Causation: A User's Guide*. Oxford University Press, Oxford. [3], [5], [18]

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, second edn. [3]

Sartorio, Carolina. 2016. *Causation & Free Will.* Oxford University Press, Oxford. [10]

Schaffer, Jonathan. 2005. "Contrastive Causation." *The Philosophical Review*, vol. 114 (3): 297–328. [4]

Thomson, Judith Jarvis. 2003. "Causation: Omissions." *Philosophy and Phenomenological Research*, vol. 66 (1): 81–103. [5]

Woodward, James. 1984. "A Theory of Singular Causal Explanation." *Erkenntnis*, vol. 21 (3): 231–262. [14]

—. 2003. *Making Things Happen: A Theory of Causal Explanation.* Oxford University Press, Oxford. [3]