A Partial Theory of Actual Causation*

Brad Weslake†

August 2, 2013
VERSION c4eb488

*I am grateful to Rachael Briggs, Kevin McCain, Michael McDermott, Peter Menzies, Marco Nathan, an audience at University of Sydney, and especially to Joe Halpern and Chris Hitchcock for their extensive comments on earlier versions. I dedicate this paper to Michael McDermott, whose seminars on counterfactuals and causation were an enormous influence on my thinking, but whose own work it has taken me many years to fully appreciate.

†Department of Philosophy
University of Rochester
Box 270078
Rochester, NY 14627-0078
bradley.weslake@rochester.edu
http://bweslake.org/

Contents

1 Introduction

In this paper I defend a partial theory of actual causation[1]. The theory is an attempt to improve on theories of actual causation developed by Hitchcock (2001), Woodward (2003, §2.7), and Halpern and Pearl (2005). These theories are all formulated in terms of causal models involving structural equations relating variables. They are subject to a number of counterexamples, which, as I will explain below, can be divided into two classes. The first class contains counterexamples which show that structurally identical causal models apply to situations which differ concerning whether a variable is an actual cause. The second class contains the remainder. My theory is partial because it only purports to address this second class of counterexamples. One way to think of a partial theory of this sort is as a sieve for eliminating all of the non-causes of an effect that can be discerned at the level of counterfactual structure. A complete theory would also involve a component that eliminates additional non-causes that can only be discerned by context-specific conditions on the values of the variables[2]. The partial theory I defend is constructed by adding an additional set of necessary conditions to the theory of Halpern and Pearl (2005).

The structure of the paper is as follows. In §2 I introduce the causal modelling framework, and provide a preliminary definition of actual cause. I then explain the various amendments to the preliminary definition proposed by Hitchcock (2001), Woodward (2003, §2.7) and Halpern and Pearl (2005). In §3 I describe the counterexamples to these theories, and show how they can be divided into the two classes above. In §4 I present my amendment to Halpern and Pearl (2005), and in §5 I show that the amendment addresses the relevant class of counterexamples. I conclude in §6.

2 Actual Causation in Causal Models

The theories of actual causation I evaluate in this paper all employ the framework of causal models[3]. In this section I will provide a minimal overview of this framework, in order to make explicit important assumptions and to provide definitions that will

[1] Actual causation is what is sometimes called "token causation", "singular causation", or "event causation". I follow the authors I here criticise in using the term made popular by Pearl (2000).

[2] An alternative way to view a partial theory of this form, recommended by Hitchcock (2007), is that it isolates one component of our thinking about actual causation, which may interact in various ways with other components of our thinking. In this case, my claim is to better have identified this component.

[3] A similar theory, developed without appeal to causal models, is defended by Yablo (2002, 2004).

be employed to formulate the theories that follow[4].

A causal model is a representational device for encoding counterfactual relationships between variables. Counterfactual relationships are represented by equations. The possible values of variables must represent entities capable of being changed by interventions, but the framework is otherwise consistent with a range of different metaphysical views concerning the nature of the causal relata. More formally, a causal model is an ordered pair $\langle \mathscr{V}, \mathscr{E} \rangle$, where $\mathscr{V}$ is a set of variables and $\mathscr{E}$ a set of equations which specify the way in which the value of a single variable on the left hand side would change as a function of the variables on the right hand side, where every variable appears on the left hand side of exactly one equation.

I will follow the usual representational conventions for variable values, so binary variables representing the occurrence or non-occurrence of an event will be assigned values 1 and 0 respectively. While in principle variables may be continuous, the examples below will always involve discrete variables with no more than three possible values. For simplicity, I will assume that all causes are represented by single variables. I will refer to a possible assignment of values to all variables in a model as a *state* of the model. I will talk freely of actual and possible variable values, changes to variable values, and states and changes of state of models. This sort of talk should be interpreted throughout as reflecting corresponding actual or possible changes in what is represented by the model. Moreover, I will assume throughout that a causal model must be veridical, in the sense that every counterfactual relationship specified by the model is true.

An equation that simply assigns a specific actual value to a variable is *exogenous*, while an equation that assigns a value as a function of other variables is *endogenous*. When displaying the equations in the examples below, I will list the exogenous and endogenous equations on separate lines. As is customary, I will assume that the equations are all deterministic, in which case the equations for a model entail the actual values of all variables in the model.

The equations are not to be interpreted symmetrically. I will use ":=" to represent an assignment of values to variables on the left hand side in the manner specified on the right hand side; inequalities ("$=$", "$\neq$", "$>$", "$<$", "$\geqslant$", "$\leqslant$") to represent functions returning 1 if the inequality is satisfied and 0 otherwise; "$\vee$" a function returning 1 if either side is 1 and 0 otherwise; "$\wedge$" a function returning 1 if both sides are 1 and 0 otherwise; and "$\neg$" a function returning 1 if input 0 and *vice versa*. So for example,

[4]For a detailed philosophical overview see Woodward (2003), and for introductions slightly more detailed than mine see Hitchcock (2001, 2007). For a detailed technical overview see Pearl (2009) and for the problem of inferring causal models from statistical data see Spirtes, Glymour, and Scheines (2000).

the equation $A := B \vee \neg D$ is to be interpreted as assigning a value of 1 to A if B is 1 or D is 0, and assigning a value of 0 otherwise. The equation also specifies whether and how A would change were the values of B or D or both to change. That is, the equations entail not just the actual values of all variables, but also the truth values of all counterfactuals concerning how the values of variables would change as a function of other variables changing values.

In particular, the equations specify the results of possible *interventions*. An intervention is an external change to the value of a variable in a model, in the sense that the values of the other variables in the model are not themselves causes or effects of the change, unless they are effects of the variable intervened on. Moreover, it is required that interventions be *surgical*, in the sense that the usual causes of the variable in question are suspended, so that the value of the variable depends only on the intervention[5].

In the literature on causation it has been common to distinguish between type-causal relations and token-causal relations. An analogous distinction can be made between between causal relations between variables, and causal relations between variable values. While the terminology is slightly misleading, I will follow Woodward (2003) and refer to causal relations between variables as *type-level* causal relations[6]. For a given causal model $\mathscr{M}$, a variable X is a (type-level) cause of a variable Y *iff* there is some state of $\mathscr{M}$ for which an intervention on X would change the value of Y. All theories of actual causation formulated in this framework agree that the actual value x of X is an actual cause for the actual value y of Y in $\mathscr{M}$ *iff* an intervention setting $X = x'$ where $x \neq x'$, with other variables in $\mathscr{M}$ held fixed by interventions at some combination of permissible possible values, would result in $Y = y'$ where $y \neq y'$. As we will see, the central difference between the theories of actual causation under examination concerns how to spell out what counts as a permissible setting of variables in this schema. Note that these are model relative definitions, which we derelativise as follows. X is a (type-level) cause of Y *simpliciter iff* there is an appropriate model in which it is so represented. Likewise, the value of X is an actual cause of Y *simpliciter iff* there is an appropriate model in which it is so represented. I will say more more about what makes a model appropriate in §4.

The theories I consider below are formulated with different vocabularies, which at times diverge even over identical concepts. Both to simplify the discussion and to make it easier to compare the theories, I employ a vocabulary which is closest to Wood-

[5]Interventions must also be statistically independent of the values of other variables in the model. For a formal definition see Woodward (2003, p. 98).

[6]For a discussion of the relationship between type-causal relations, token-causal relations, and causal relations between variables, see Hausman (2005).

ward (2003). While I provide references where appropriate, the precise formulations I give are sometimes simplified or expanded, and sometimes make use of definitions introduced in this paper.

2.1 A Preliminary Theory

It will help to motivate the theories I examine below to consider a preliminary theory of actual causation that can be defined in the causal modelling framework. We start with the (type-level) notion of a *direct cause* (Woodward 2003, p. 55):

> (DC) X is a direct cause of Y in model $\mathscr{M}$ *iff* there is a possible intervention on X that would change Y when all other variables in $\mathscr{M}$ besides X and Y are held fixed at some combination of values by interventions.

It is a necessary and sufficient condition for X to be a direct cause of Y in $\mathscr{M}$ that X appear on the right hand side of the equation for Y in $\mathscr{M}$. Next we need the (type-level) notions of a *directed graph* and a *directed path* (*ibid*, p. 42):

> (G) A *directed graph* for $\mathscr{M}$ is an ordered pair $\langle V, E\rangle$ where V is the set of variables in $\mathscr{M}$ and E a set of ordered pairs of elements of V (hereafter *directed edges*), where there is a directed edge from X to Y *iff* X directly causes Y in $\mathscr{M}$.
>
> (P) A sequence of variables $\{V_1 \ldots V_n\}$ is a *directed path* from V_1 to V_n in $\mathscr{M}$ *iff* for all $i (1 \leqslant i < n)$ there is a directed edge from V_i to V_{i+1} in the directed graph for $\mathscr{M}$.

From here on, *path* should be read as equivalent to *directed path*. Paths between variables when a model is not too large can easily be seen by constructing a diagram with the same structure as the associated directed graph, which I will provide for all of the examples below. In these diagrams, labelled circles correspond to variables and arrows correspond to directed edges (direct causal relations). Next we define the (type-level) notion of a *contributing cause* (*ibid*, p. 59):

> (CC) X is a *contributing cause* of Y in model $\mathscr{M}$ *iff* for some path P from X to Y in $\mathscr{M}$, there is an intervention on X that will change Y when all variables in $\mathscr{M}$ not on P are held fixed at some combination of values by interventions.

Note some consequences of these definitions. First, as will become clear from the theories of actual causation below, if $X = x$ is an actual cause of $Y = y$ in $\mathscr{M}$ then

X is a contributing cause of Y in $\mathscr{M}$. Second, notice that each of these definitions is relativised to a causal model. While it is acceptable to leave (DC) and (P) as model-relative notions, the corresponding de-relativised definition for (CC) is as follows[7]: X is a *contributing cause* of Y *simpliciter iff* there exists an appropriate model in which X is a contributing cause of Y. Likewise, for every candidate model-relative definition of actual cause below, I will assume that the proposal implicitly includes the claim that $X = x$ is an *actual cause* of $Y = y$ *simpliciter iff* there exists an appropriate model in which $X = x$ is an actual cause of $Y = y$.

All of the theories of actual causation I consider in this section can be formulated as instances of the following schema:

(AC) $X = x$ is an *actual cause* of $Y = y$ relative to model $\mathscr{M}$ *iff*:

> (ACT) The actual value of $X = x$ and the actual value of $Y = y$.
>
> (PATH) There exists a path P_i from X to Y in $\mathscr{M}$ for which an intervention on X would change the value of Y, when all variables $V_1 \ldots V_n$ in $\mathscr{M}$ that are not on P_i are held fixed at some combination of values satisfying <*conditions specifying permissible values* $v_1 \ldots v_n$ *for* $V_1 \ldots V_n$>.

The conditions on permissible values can be thought of as specifying the set of possible values of the off-path variables relative to which an intervention constitutes a test for actual causation along that path. All theories I consider in detail below agree that *one* such permissible set is that in which all off-path variables have their actual values. So they all agree that a sufficient condition for $X = x$ to be an actual cause of $Y = y$ is for there to be a path from X to Y such that holding all off-path variables fixed at their actual values, there is an intervention setting $X = x'$ where $x \neq x'$ that would result in $Y = y'$ where $y \neq y'$[8].

The preliminary definition of actual causation with which I will begin can be specified by turning this sufficient condition into a necessary and sufficient condition, by imposing the following condition on permissible values:

(PREP) $v_1 \ldots v_n$ are the actual values of $V_1 \ldots V_n$.

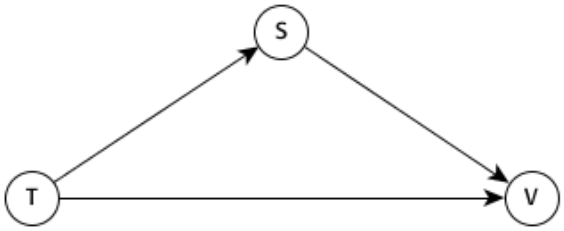
[7]Here I follow Hitchcock (2007, p. 503) and Woodward (2008, p. 209).

[8]As a referee pointed out to me, not all theories formulated in the framework of causal models agree. The difference concerns the way in which these theories treat the notion of a relevant possibility, to be introduced in §4: some theories disallow values reflecting irrelevant possibilities from appearing in a model, while others impose further restrictions on the conditions sufficient for actual causation.

Backup
If Trainee shoots his gun (T), the bullet will hit Victim (V). If Trainee does not shoot, Supervisor will shoot and hit Victim herself (S). In fact, Trainee shoots and hits Victim while Supervisor stands by.

$$T := 1 \tag{EX}$$

$$S := \neg T, V := T \vee S \tag{ED}$$



Example 1: *Directed Graph for Backup*

(PRE$_P$) provides us with our first account of which possible values for off-path variables it is permissible to hold fixed when testing for actual causation. According to (PRE$_P$), we must hold fixed *all* off-path variables at their *actual* values. The theory constructed by plugging (PRE$_P$) into (AC) I will call (PRE)[9].

An example where (PRE) fares well is **Backup** (§1, p. 8), taken from Hitchcock (2001, p. 276). In this example, there is no intervention on T that would change V, with *no* other variables held fixed. But if we hold S fixed at the actual value 0, then an intervention setting T to 0 would change V to 0. So according to (PRE), $T = 1$ is an actual cause of $V = 1$. This is the right result, and suggests that the preliminary definition is on the right track. However, a class of cases for which (PRE) fares poorly are those involving symmetric overdetermination. Take for instance the well known example **Window** (§2, p. 9), taken from Hall (2004b, p. 278). In this example, (PRE) mistakenly entails that neither $B = 1$ nor $S = 1$ is an actual cause of $W = 1$.

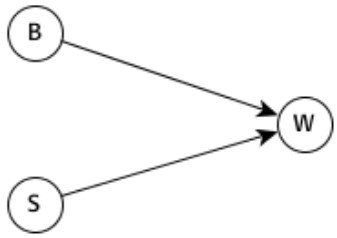
The various amendments to (PRE) I consider in the following subsections are all motivated by the thought that the reason (PRE) fails in cases of this sort is that the causal influence on one path is masked by the causal influence on another path. The amendment motivated by this thought is to allow that it is permissible to hold fixed

[9](PRE) is equivalent (modulo some irrelevant differences) to Woodward's (AC) (2003, p. 77), and the definition of causation defined in terms of "Act" in Hitchcock (2001, pp. 286–287).

Window
Billy ($B = 1$) and Suzy ($S = 1$) both throw rocks at a window, each with sufficient force to shatter it. The rocks strike the window at exactly the same time. The window breaks ($W = 1$).

$$S := 1, S := 1 \qquad \text{(EX)}$$

$$W := B \vee S \qquad \text{(ED)}$$



Example 2: *Directed Graph for Window*

some off-path variables to *some* non-actual values when testing for actual causation. Which off-path variables, and which non-actual values? Intuitively, just those variables which are masking the influence of the variable in question along the path in question, and just those non-actual values which remove the mask. As we will see, no account has yet hit on a condition that successfully captures this intuition.

2.2 Woodward

Woodward's proposed amendment to (PREP) is the following condition:

> (WP) No intervention setting $V_1 \ldots V_n$ to $v_1 \ldots v_n$ while holding the actual value of X fixed would result in a change to the actual value of Y.

(WP) provides us with our second account of which possible values for off-path variables it is permissible to hold fixed when testing for actual causation. According to (WP), we may hold fixed all off-path variables at any combination of their actual or non-actual values, so long as interventions to those values, holding the candidate cause fixed, makes no difference to the candidate effect.

The theory constructed by plugging (WP) into (AC) I will call (W)[10]. Note that since holding fixed all off-path variables at their *actual* values, and holding the candidate

[10](W) is equivalent (modulo some irrelevant differences) to Woodward's (AC*) (2003, p. 84).

cause fixed, makes no difference to the candidate effect, every actual cause identified by (PRE) is also identified by (W), a relationship I will represent by writing (PRE) < (W). Since (PRE) < (W), (W) correctly handles **Backup**.

(W) improves upon (PRE) with respect to cases of symmetric overdetermination. Return to **Window**. All possible values of S satisfy (W_P) for the path $\{B \rightarrow W\}$, since no intervention on S while holding B fixed would result in a change to W. So we are permitted to test for the efficacy of B by setting $S = 0$. But if $S = 0$, setting $B = 0$ would result in $W = 0$. So $B = 1$ is an actual cause of $W = 1$ according to (W). Turning to $\{S \rightarrow V\}$, we see that the situation is symmetrical. So $B = 1$ is also an actual cause of $W = 1$ according to (W). This is the right result, so we are still on the right track.

Nonetheless, (W) does not fare well with cases involving late preemption. Consider the well known example **Bottle** (§3, p. 11) taken from Hall (2004b, p. 235). In this example, $SH = 0$ satisfies (W_P) for path $\{BT \rightarrow BH \rightarrow BS\}$, for if Billy had still thrown but Suzy's rock not hit the bottle, the bottle still would have smashed. But if we hold fixed $SH = 0$, then setting $BT = 0$ would result in $BS = 0$. So (W) mistakenly counts Billy's throw an actual cause.

Intuitively, what has gone wrong is that in permitting SH to change value, in order to remove the potential for masking of Suzy's throw, our condition inadvertently allowed a change to BH, which in turn enabled BT to count as an actual cause of BS. That is, in unmasking influence along path $\{ST \rightarrow SH \rightarrow BS\}$, we illegitimately activated influence along path $\{BT \rightarrow BH \rightarrow BS\}$.

2.3 Hitchcock

The theory of actual cause provided by Hitchcock (2001, p. 290) does not have this problem[11]. Hitchcock's proposed amendment to (PRE_P) is the following condition:

> (H_P) No intervention setting $V_1 \ldots V_n$ to $v_1 \ldots v_n$ while holding the actual value of X fixed would result in a change to the actual values of any variables on P_i.

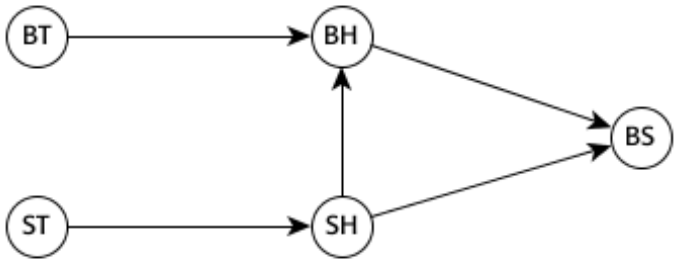
(H_P) provides us with our third account of which possible values for off-path variables it is permissible to hold fixed when testing for actual causation. Unlike (W_P), (H_P) requires that the actual values of variables along the path from X to Y be preserved by interventions on $V_1 \ldots V_n$. The theory constructed by plugging (H_P) into (AC) I will call (H). It is easy to see that (PRE) < (H) < (W).

[11] Woodward (2003, p. 83) says that his proposal is "due to" Hitchcock (2001) and (an early version of) Halpern and Pearl (2001). As we will see, none of these proposals are equivalent.

Bottle
Suzy (ST) and Billy (BT) both throw rocks at a bottle, Suzy's rock arrives first and hits the bottle (SH), the bottle shatters (BS), Billy's arrives second and does not hit the bottle (BH). Both throws are accurate, Billy's would have shattered the bottle if Suzy's had not.

$$\mathsf{BT} := 1, \mathsf{ST} := 1 \qquad \text{(EX)}$$

$$\mathsf{SH} := \mathsf{ST}, \mathsf{BH} := \mathsf{BT} \wedge \neg \mathsf{SH}, \mathsf{BS} := \mathsf{BH} \vee \mathsf{SH} \qquad \text{(ED)}$$



Example 3: *Directed Graph for Bottle*

(H) improves on (W) with respect to cases involving late preemption. Returning to **Bottle**, SH = 0 does *not* satisfy (HP) for path {BT → BH → BS}, for if Billy had still thrown and Suzy's rock not hit the bottle, Billy's rock would have hit the bottle. So the actual value of variable BH on path {BT → BH → BS} would be changed, and hence SH = 0 does not satisfy (HP). So (H) correctly rules that Billy's throw is not a cause. Likewise, (H) correctly rules that Suzy's throw is a cause, and retains all the successes of (W) for the preceding examples[12]. So we remain on the right track.

Nonetheless, it turns out that while (W) is too permissive, (H) is not permissive enough. Consider the example **Vote Machine** (§4, p. 13), taken from Halpern and Pearl (2005, p. 881). (H) incorrectly rules that neither V1 nor V2 are actual causes of P, since there is no intervention on either that does not change M, which is on both paths to P.

Intuitively, what has gone wrong is that in disallowing *any* changes to variables along a path when testing for actual causation, we ruled out certain *innocuous* interactions between paths. As **Vote Machine** shows, not all interactions between paths illegitimately activate non-causes, and in some cases allowing variables on a path to *innocuously* change value may be essential to revealing influence along that path. The challenge is to specify what counts as an innocuous change.

2.4 Halpern and Pearl

Halpern and Pearl present three different theories of actual causation over the course of the two different published versions of their paper (Halpern and Pearl, 2001, 2005). In this section I will present the first two theories, as the third is motivated in part by the class of counterexamples I have set aside and inherits the problems with the first two on which I will focus[13]. The first proposal is as follows (Halpern and Pearl 2001)[14]:

> (HPI$_P$) No intervention setting $V_1 \ldots V_n$ to $v_1 \ldots v_n$ while holding the actual value of X fixed would result in a change to the actual value of Y, even if an

[12] I leave the details as an exercise, though since the only difference between (W) and (H) concerns a condition that can only be satisfied when there are overlapping paths, their equivalence can often by seen by simply inspecting the associated directed graphs.

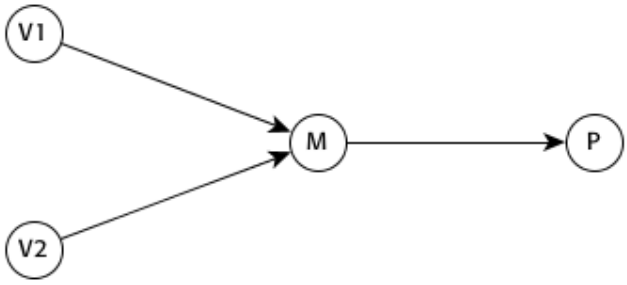
[13] See Halpern (2008) for reservations about this third theory with respect to the examples I have set aside.

[14] As Joe Halpern pointed out to me, my formulations of these conditions are not strictly equivalent to the conditions Halpern and Pearl introduce, since mine are formulated in terms of a single path between X and Y and theirs are not. I use the present formulation because it is simpler, and gives the same result for all of the examples I discuss.

Vote Machine
Two votes ($V1$, $V2$) are cast for a measure. The votes are summed by a machine (M) and the measure passes (P) *iff* it receives at least one vote.

$$V1 := 1, V2 := 1 \qquad \text{(EX)}$$

$$M := V1 + V2, P := (M \geqslant 1) \qquad \text{(ED)}$$



Example 4: *Directed Graph for Vote Machine*

arbitrary subset of the variables in P_i were set to their actual values by interventions.

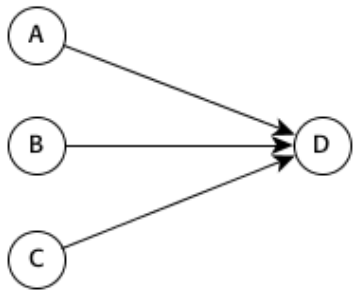
(HPI$_P$) provides us with our fourth account of which possible values for off-path variables it is permissible to hold fixed when testing for actual causation. Unlike (H$_P$), (HPI$_P$) allows the actual values of variables along the path from X to Y to be changed by interventions on $V_1 \ldots V_n$, but unlike (W$_P$), they may only be changed if setting (arbitrary subsets of) them back to their actual values would not make a difference to the candidate effect. The theory constructed by plugging (HPI$_P$) into (AC) I will call (HPI). Again, it is easy to see that (PRE) < (H) < (HPI) < (W).

(HPI) improves on (H) with respect to **Vote Machine**. Setting $V2 = 0$ does not satisfy (H$_P$) for path $\{V1 \rightarrow M \rightarrow P\}$, since doing so would change M. But it does satisfy (HPI$_P$), for although setting $V2 = 0$ would result in $M = 1$, subsequently setting M back to the actual value $M = 2$ would not change P. So (HPI) successfully identifies $V1$ and $V2$ as actual causes of P. (HPI) also gets the right result for **Bottle**. In contrast to (W$_P$), $SH = 0$ does not satisfy (HPI$_P$) for path $\{BT \rightarrow BH \rightarrow BS\}$. Setting $SH = 0$ would result in $BH = 1$, and subsequently returning BH to the actual value $BH = 0$ would change BS. So we are on the right track, having found a condition that promises to discriminate between innocuous (as in **Vote Machine**) and illegitimate (as in **Bottle**) changes to on-path variables.

Loader
A firing squad consists of shooters B and C. It is A's job to load B's gun, C loads and fires his own gun. On a given day, A loads B's gun. When the time comes, only C shoots the prisoner.

$$A := 1, B := 0, C := 1 \qquad \text{(EX)}$$

$$D := (A \wedge B) \vee C \qquad \text{(ED)}$$



Example 5: *Directed Graph for Loader*

As it turns out, (HPI) is again too permissive, continuing our trend of oscillating between overly permissive and overly restrictive theories. Consider example **Loader** (§5, p. 14), taken from Hopkins and Pearl (2003). In this example, $\{B = 1, C = 0\}$ satisfies (H_P) for path $\{A \rightarrow D\}$, for setting $\{B = 1, C = 0\}$ would not change D, and there is no path interaction to be ruled out as illegitimate. So (H) generates the mistaken verdict that $A = 1$ is an actual cause of $D = 1$. And since (H) < (HPI) < (W), so do (HPI) and (W).

Halpern and Pearl (2005) respond to this example with the final proposal I will consider in this section. Their revised condition is (p. 853):

> ($HP2_P$) No intervention setting an arbitrary subset of $V_1 \ldots V_n$ to $v_1 \ldots v_n$, while holding the actual value of X fixed, would result in a change to the actual value of Y, even if an arbitrary subset of the variables in P_i were set to their actual values by interventions.

($HP2_P$) provides us with our fifth account of which possible values for off-path variables it is permissible to hold fixed when testing for actual causation. ($HP2_P$) simply adds to (HPI_P) the requirement that Y not be changed by setting arbitrary subsets of variables $V_1 \ldots V_n$ to values $v_1 \ldots v_n$, even if an arbitrary subset of the variables in P_i

were set to their actual values by interventions. The theory constructed by plugging (HP2$_P$) into (AC) I will call (HP2). Since we have simply imposed another necessary condition, it is easy to see that (HP2) < (HPI) < (W). It is also easy to see that (PRE) < (HP2).

There is no such simple inclusion relationship between (HP2) and (H), as reflection on two of our examples illustrates. (HP2) fares well for **Loader**. $\{B = 1, C = 0\}$, which satisfies (H$_P$) for path $\{A \rightarrow D\}$ and thereby generates the mistaken verdict, is ruled out by (HP2$_P$). For an intervention setting $C = 0$, which is a subset of that set, would change D. So (H) $\not<$ (HP2). Moreover, as can be seen by reconsidering **Vote Machine**, (HP2) $\not<$ (H).

However, note that when the set of off-path variable values satisfying (HPI$_P$) only contains a single variable set to a non-actual value, (HP2$_P$) will automatically be satisfied. When this condition obtains I will say that the off-path variable values satisfying (HPI$_P$) are *simple*. And when (HPI) is equivalent to (HP2), I will say that (HPI)≡(HP2). When the off-path variable values are simple, then, (PRE) < (H) < (HPI)≡(HP2) < (W).

As I will demonstrate below, (HP2) is not the correct way to respond to **Loader**. For both (HPI) and (HP2) face a set of counterexamples that are best addressed by an alternative amendment to (HPI)—or so I will argue[15].

3 Counterexamples

In this section I describe the counterexamples to (HPI) and (HP2) that will motivate my proposed amendment. I note at the outset that not all of these counterexamples are without controversy, and I will provide references to some of those who have different intuitions, or who believe we should mistrust our intuitions in the relevant cases. But I believe that they collectively motivate a theory that is capable of addressing them. I finish the section by providing some of the counterexamples that my theory is not designed to address.

3.1 Switching

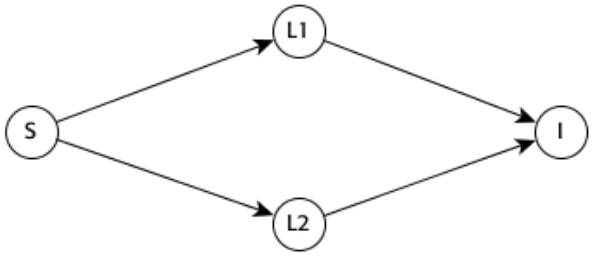
Consider example **Switch** (§6, p. 16), taken from Pearl (2000, p. 324). According to (PRE), $S = 1$ is an actual cause of $I = 1$, for holding $L1$ fixed at the actual value $L1 = 0$, an intervention setting $L2 = 0$ would result in $I = 0$. Since every off-path variable

[15]Chris Hitchcock (personal communication) and a referee have suggested that **Loader** should be treated as a preemption case, with additional variables representing whether the guns fire. This is a better way to handle this particular case, but will not help for the cases to be discussed below.

Switch
A two-state switch is wired to two lamps. If the switch is in one state ($S = 0$), only lamp one is activated ($L1 = 1$), and if it is in the other state ($S = 1$) only lamp two is activated ($L2 = 1$). In fact, lamp two is activated and the room is illuminated ($I = 1$).

$$S := 1 \qquad \text{(EX)}$$

$$L1 := \neg S, L2 := S, I := L1 \vee L2 \qquad \text{(ED)}$$



Example 6: *Directed Graph for Switch*

set is a singleton, (PRE) < (H) < (HP1)≡(HP2) < (W), and so this is also the verdict of every other theory we have considered. But this is a mistake, for the position of the switch is not an actual cause of the room being illuminated.

An example with a similar structure is **Shock** (§7, p. 17), taken from McDermott (1995, p. 532). According to (PRE), $A = 1$ is an actual cause of $C = 1$, for holding B fixed at the actual value $B = 1$, an intervention setting $A = 0$ would result in $C = 0$. Since every off-path variable set is a singleton, (PRE) < (H) < (HP1)≡(HP2) < (W), and so this is also the verdict of every other theory we have considered. But this is a mistake, for A flipping her switch is not an actual cause of C being shocked.

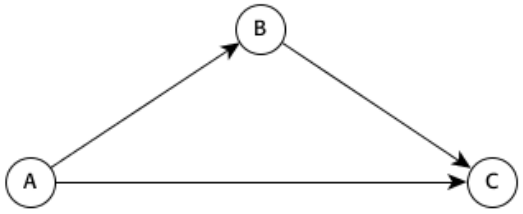
These claims are not uncontroversial. Both Hall (2000)[16] and Pearl (2000, §10.3.4), for different reasons, are content to count switches as causes. But I count it a virtue of a theory if it can return the majority verdict.

[16]Though Hall (2007, p. 118, fn. 9) reflects: "I labored mightily to have the contrary intuition, in order to preserve the transitivity of causation. I now think that was probably a mistake".

Shock
Two two-state switches are wired to an electrode. The switches are controlled by A and B respectively, and the electrode is attached to C. A has the first option to flip her switch ($A = 1$). B has the second option to flip her switch ($B = 1$). The electrode is activated and shocks C ($C = 1$) *iff* both switches are in the same position. B wants to shock C, and so flips her switch *iff* A does.

$$A := 1 \qquad \text{(EX)}$$

$$B := A, C := (A = B) \qquad \text{(ED)}$$



Example 7: *Directed Graph for Shock*

3.2 Trumping

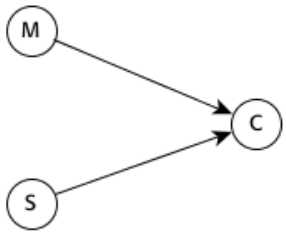
Consider example **Command** (§8, p. 18), adapted from Schaffer (2000, p. 175). According to (PRE), $S = 1$ is not an actual cause of $C = 1$. But according to (H), $S = 1$ is an actual cause of $C = 1$. For $M = 0$ satisfies (H_P) for path $\{S \rightarrow C\}$, and when $M = 0$ setting $S = 0$ would result in $C = 0$. Since every off-path variable set is a singleton, (H) < (HP1)≡(HP2) < (W), and so this is also the verdict of the remaining theories we have considered. But this is a mistake, for the sergeant shouting 'Charge!' is not an actual cause of the corporal charging.

This example is probably the most controversial of those I will take into account, as there has been a surprising amount of debate concerning whether S is a cause of C, and more generally whether trumping causation is a species of preemption or overdetermination. On the side of preemption we have Schaffer (2000), Lewis (2000), and the "preanalytic assessment" of Woodward (2003, p. 382 fn. 44), who also believes this is the majority assessment. On the side of overdetermination we have McDermott (2002), Halpern and Pearl (2005) and Hitchcock (2011). While I place the least weight on this example, I still count it a virtue of a theory that it can deliver what I concur with Woodward in believing is the majority assessment.

Command
Major (M) and and sergeant (S) stand before corporal, and both shout 'Charge!' ($M = 1, S = 1$). The corporal charges ($C = 1$). Orders from higher-ranking soldiers trump those of lower rank, so if the major had shouted 'Halt' ($M = 2$) the corporal would not have charged.

$$M := 1, S := 1 \qquad \text{(EX)}$$

$$C := (M = 1) \vee (S \wedge M \neq 2) \qquad \text{(ED)}$$



Example 8: *Directed Graph for Command*

3.3 Combination Lamp

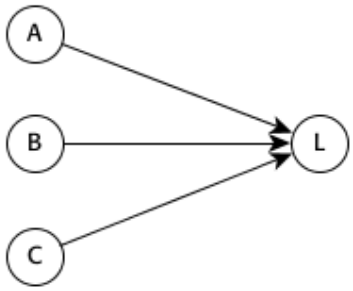
Consider next an example I call **Combination Lamp** (§9, p. 19). In this example, $\{B = 1, C = -1\}$ satisfies (H_P) for path $\{A \rightarrow L\}$, and when $B = 1$ and $C = -1$ setting $A = 0$ would result in $L = 0$. So $A = 1$ is an actual cause of $L = 1$ according to (H). Moreover, since this off-path variable set only involves a single change to a non-actual value, (H) < (HP1)≡(HP2) < (W), and so all of these theories deliver the same result. But this is a mistake, for switch A being in position 1 is not an actual cause of the lamp being on.

We are now in a position to diagnose the problem with (HP1), and to see why (HP2) is an inadequate response to **Loader**. There is a single feature of **Loader** and **Combination Lamp** that leads to the problem with (HP1). In each of these examples, there is some variable $X = x$ that is not an actual cause of $Y = y$, but which would have been an actual cause if certain other variables V_i had taken different values. Moreover, in each of these examples, variables V_i can be changed to these values in a way that satisfies (H_P). This reveals that it is an entirely incidental feature of **Loader** that every set of variables V_i that satisfies (H_P) contains a subset that does not.

Since **Loader** and **Combination Lamp** are instances of the same problem, (HP2) is not on the right track. Instead, we need a condition that rules out this means of

Combination Lamp
A lamp (L) is controlled by three switches (A, B and C), each of which has three possible positions $(-1, 0, 1)$. The lamp switches on *iff* two or more switches are in the same position. In fact, the switches are in positions $A = 1$, $B = -1$ and $C = -1$.

$$A := 1, B := -1, C := -1 \qquad \text{(EX)}$$
$$L := (A = B) \vee (B = C) \vee (A = C) \qquad \text{(ED)}$$



Example 9: *Directed Graph for Combination Lamp*

illegimately activating a non-cause[17].

3.4 Non-Structural Counterexamples

In this subsection I describe two counterexamples that will help to clarify the sense in which the theory I will propose in the following section is partial. The counterexamples discussed above are all structural in character, in the sense that the non-causes in these examples are also non-causes in any example appropriately modelled by the same structural equations[18]. But not all counterexamples are of this type.

Consider example **Careful Antidote** (§10, p. 20), taken from Hiddleston (2005, p. 32)[19]. In this example, $A = 0$ is an actual cause of $D = 0$ according to (H). Since

[17]It is curious that Halpern and Pearl (2005) did not recognise the possibility of examples such as **Combination Lamp**, for Hopkins and Pearl (2003) proved a result ("Theorem 3") that entails their existence. I believe that **Combination Lamp** exhibits the simplest possible structure of this form. As an autobiographical matter, I first saw these possibilities by reflecting on an example due to Hall (2004a, p. 273), and then later discovered Hopkins and Pearl (2003).

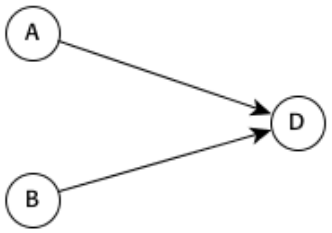
[18]I do not claim that this is obvious, but I will not argue for it here.

[19]A similar example is given by Hall (2007, §3.2)

Careful Antidote
Assassin is in possession of a lethal poison, but has a last-minute change of heart and refrains from putting it in Victim's coffee ($A = 0$). Bodyguard puts antidote in the coffee ($B = 1$), which would have neutralized the poison. Victim drinks the coffee and survives ($D = 0$).

$$A := 0, B := 1 \qquad \text{(EX)}$$

$$D := A \wedge \neg B \qquad \text{(ED)}$$



Example 10: *Directed Graph for Careful Antidote*

every off-path variable set is a singleton, (H) < (HP1)≡(HP2) < (W), and so this is also the verdict of these theories. But this is a mistake, for delivering the antidote is not an actual cause of survival.

The important thing to notice about the causal model for this example is that it is structurally identical to simple cases of symmetric overdetermination such as **Window**[20]. In simple cases of symmetric overdetermination, the effect occurs *iff* one or both of the causes occur. Likewise, in **Careful Antidote**, Victim survives *iff* one or both of the non-poisoning and antidote delivery occur. So if we wish to treat these cases differently with a theory of actual causation, the theory will need to add resources that go beyond the counterfactual structure of the case[21].

Similarly, consider example **Careful Poisoning** (§II, p. 22), taken from Hitchcock (2007, pp. 519*ff*)[22]. According to (PRE), $A = 1$ is an actual cause of $D = 0$, for holding B fixed at the actual value $B = 1$, an intervention setting $A = 0$ would result in $D = 1$.

[20]This is pointed out by Halpern (2008, §4) and Halpern and Hitchcock (2010, p. 400).

[21]Chris Hitchcock (personal communication) suggests that **Careful Antidote** should be treated as a preemption case, with an additional variable representing whether the poison is neutralised. If so, then we can set this example aside and focus on the next.

[22]Hitchcock credits McDermott (personal communication) and Björnsson (2007) for identifying cases of this sort.

Since every off-path variable set is a singleton, (PRE) < (H) < (HP1)≡(HP2) < (W), and so this is also the verdict of every other theory we have considered. But this is a mistake, for delivering the antidote is not an actual cause of survival[23].

Again, the important thing to notice about the causal model for this example is that it is structurally identical to **Backup** (§1, p. 8). To see this, let B = 0 represent poisoning the coffee and B = 1 represent not poisoning the coffee, let D = 0 represent dying and D = 1 represent surviving, and rewrite the corresponding equations accordingly[24]. So again, if we wish to treat these cases differently with a theory of actual causation, the theory will need to add resources that go beyond the counterfactual structure of the case.

It is non-structural counterexamples of this sort that I am here setting to one side. My aim is to provide a partial theory that eliminates all of the non-causes of an effect that can be discerned at the level of counterfactual structure, and to do this the theory must rule against switches, against the trumped, and against the switch in **Combination Lamp**. But it need not rule against the cases in this subsection and others in the same class[25].

4 A Partial Theory of Actual Causation

In this section I present a partial theory of actual causation. The theory combines a number of different elements, many of which have received expression somewhere or other in the literature. But they have not yet been combined, or expressed within the framework of causal models, in the way required to handle all of the examples above. In this section I will formulate the theory, and in the next I will apply it to the examples[26].

[23] This claim is not uncontroversial. Hitchcock (2007, *loc cit*) reports that some of his respondents have the reverse intuition, and many have no clear intuition.

[24] This contravenes the representational convention we have been using, according to which we use 1 to represent the occurrence of an event and 0 to represent the non-occurrence of an event, but those conventions are of course arbitrary.

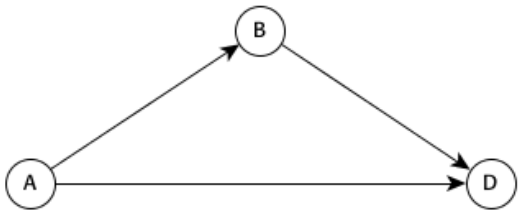
[25] The most popular way to handle these cases is to appeal to a contextually determined distinction between default and deviant variable values. See Menzies (2004, 2007), Hitchcock (2007), Hall (2007), Halpern (2008), Hitchcock and Knobe (2009) and Halpern and Hitchcock (2010). Though broadly sceptical of counterfactual accounts of causation, similar ideas are developed by Maudlin (2004).

[26] In the causal modelling literature, the theories that are closest to mine are those formulated in terms of "contributory cause" in Park (2003) and "actual causation" in Baldwin and Neufeld (2004). The Pearl (2000) theory in terms of "causal beam" is more distantly related, as it delivers the wrong result in **Vote Machine**. In the philosophy literature, the theories that are closest are McDermott (1995, 2002), and more distantly Strevens (2007; 2008, Chapter 6).

Careful Poisoning
Assistant Bodyguard puts a harmless antidote in Victim's coffee ($A = 1$). Buddy then poisons the coffee ($B = 1$), using a poison that is normally lethal, but which is countered by the antidote. Buddy would not have poisoned the coffee if Assistant had not administered the antidote first. Victim drinks the coffee and survives ($D = 0$).

$$A := 1 \qquad \text{(EX)}$$

$$B := A, D := \neg A \wedge B \qquad \text{(ED)}$$



Example 11: *Directed Graph for Careful Poisoning*

First I will introduce the notion of a *weakly sufficient* set of variable values for the actual value of a variable in a state of a model. Let us say that a set of variable values $\{X_1 = x_1 \ldots X_n = x_n\}$ is *weakly sufficient* for $Y = y$ where $X_i \neq Y$ in a state of $\mathscr{M}$ *iff* every variable has the actual value specified by the set, and there is no possible combination of values for the remaining variables in $\mathscr{M}$ that would change the value of $Y = y$ if $\{X_1 \ldots X_n\}$ were held fixed at $\{x_1 \ldots x_n\}$ by interventions[27]. Less formally, variable values are weakly sufficient for another variable value when there are no interventions elsewhere in the model that would change that variable value.

Second I will introduce the notion of a *minimally weakly sufficient* (hereafter simply *minimal*) set of variable values for the actual value of a variable in a state of a model. Let us say that a set of variable values $\{X_1 = x_1 \ldots X_n = x_n\}$ is *minimal* for $Y = y$ in a state of $\mathscr{M}$ *iff* $\{X_1 = x_1 \ldots X_n = x_n\}$ is weakly sufficient for $Y = y$, and no proper subset of $\{X_1 = x_1 \ldots X_n = x_n\}$ is weakly sufficient for $Y = y$.

[27] Pearl (2000, §10.2) calls this condition *sustenance* and Halpern and Pearl (2005, p. 855) say that when it holds $\{X_1 = x_1 \ldots X_n = x_n\}$ *strongly causes* $Y = y$. McDermott's notion of "sufficient condition" (1995, p. 533; 2002, pp. 96–97) is also close, though McDermott's definition makes the assumption that we are dealing with binary events. I use the term "weak sufficiency" for consistency with other work in which I employ a stronger notion (Weslake ms).

Only values of direct causes of Y will be members of a minimal set for $Y = y$, so third I will introduce the notions of a *net* and a *strand*. Let us say that a set of variable values N is a *net* for $Y = y$ in a state of $\mathscr{M}$ *iff* there is a partition of N into sets $N_1 \ldots N_n$ such that N_1 is a minimal set for $Y = y$, and N_{i+1} is the union of a set of minimal sets, one for each element of N_i, for all $i(1 \leqslant i < n)$. And let us say that a sequence of variable values $\{V_1 = v_1 \ldots V_n = v_n\}$ is a *strand* for $V_n = v_n$ in a state of $\mathscr{M}$ *iff* V_i is a direct cause of V_{i+1}, and $V_i = v_i$ is part of a minimal set for $V_{i+1} = v_{i+1}$, for all $i(1 \leqslant i < n)$. Less formally, a net for a variable value is constructed by taking a minimal set N_1 for that value, optionally adding a minimal set for each element of N_1, and so on. And a strand is a sequence of variable values along a path, each of which is part of a minimal set of variable values for the next.

We are now in a position to introduce the partial theory of actual causation:

(PART) $X = x$ is an *actual cause* of $Y = y$ relative to model $\mathscr{M}$ *only if*:

(ACT) The actual value of $X = x$ and the actual value of $Y = y$.

(STRAND) $X = x$ is on a strand for $Y = y$ in the actual state of $\mathscr{M}$.

(PATH) There exists a path P_i from X to Y in $\mathscr{M}$ for which an intervention setting $X = x'$ would result in $Y = y'$, when all variables $V_1 \ldots V_n$ in $\mathscr{M}$ that are not on this path are held fixed at some combination of values satisfying (HPI_P):

(HPI_P) No intervention setting $V_1 \ldots V_n$ to $v_1 \ldots v_n$ while holding the actual value of X fixed would result in a change to the actual value of Y, even if an arbitrary subset of the variables in P_i were set to their actual values by interventions.

(DIF) In the state of $\mathscr{M}$ produced by an intervention setting $X = x'$, it is not the case that the preceding conditions are satisfied for $X = x'$ with respect to $Y = y$.

The basic conception of causation that (PART) enshrines is one in which a cause is a difference-making part of a chain of minimally sufficient conditions for an effect. (ACT) requires that both cause and effect actually occur. (STRAND) requires that the cause be part of a chain of minimally sufficient conditions. (PATH) and (DIF) express the requirements for the cause to make a difference to the effect. (PATH) can be thought of as providing a recipe for disabling background conditions that mask the difference made by the cause, and requiring that with the relevant background conditions disabled, the cause must make a difference to the effect in the sense that if

the cause had not occurred the effect would not have occurred. (DIF) requires further that the cause made a difference in a different sense, namely that the alternative to the cause would not have been a cause of the same effect[28].

(PART) has an obvious affinity with the well known proposal of Mackie (1974) that a cause is an insufficient but nonredundant part of an unnecessary but sufficient condition for the occurrence of an effect, or for short an (INUS) condition for the effect. While I will not engage in a detailed comparison of the present approach with Mackie, it is worth noting how (PART) does not suffer the two main problems with the bare (INUS) theory. These problems are that effects are (INUS) conditions for causes, and effects of a common cause are (INUS) conditions for each other. In (PART), these possibilities are ruled out by the structural equations themselves. Since there are no paths from effects to causes or (in general) between effects of a common cause, these problems do not arise for (PART)[29].

As with the other theories we have considered, I will take (PART) to include the claim that $X = x$ is an actual cause of $Y = y$ *simpliciter iff* there exists an appropriate model in which $X = x$ is an actual cause of $Y = y$. So far I have not made appeal to any particular conditions on the appropriateness of models. This is in part because what counts as an appropriate model is a highly contextual matter that deserves far more treatment than I can provide here, and in part because a complete account of model appropriateness is best pursued in tandem with the part of a complete theory of actual causation that I have set aside[30]. However, I will be appealing at several points below to the following principle governing model appropriateness. When we make causal judgements about a situation, we treat some features of the situation as fixed and some as potentially variable. Regarding those features we treat as potentially variable, we treat some alternative possibilities as relevant and some as irrelevant. Often, these judgements of relevance track the possibilities that fall beneath some contextually variable threshold of likelihood. The principle is this:

> *Relevance* An appropriate model should include values for variables that reflect all and only the relevant possibilities.

[28](DIF) is similar, but not identical, to a principle that Sartorio (2005) calls the *Causes as Difference-Makers* principle. Sartorio argues that this principle fits well with the connection between causation and moral responsibility, and her arguments also apply to (DIF).

[29]Similarly, McDermott (1995, 2002) defines sufficiency in terms of counterfactuals, which allows his account to retain the benefits of Lewis (1973) with respect to Mackie; and Strevens (2007, §5; 2008) takes facts about causal connection as primitive and then develops an account of actual causation using (INUS) conditions that builds on these facts.

[30]For a useful survey of principles governing appropriate models see Halpern and Hitchcock (2010).

Clearly the principle as stated requires sharpening and clarification. But this loose and uninformative formulation will be sufficient for my purposes below, and I proceed under the assumption that a full account of model appropriateness will deliver a more precisely formulated version[31].

5 Counterexamples Revisited

(PART) delivers the correct verdict in every example from §2[32]. In this section I show how (PART) improves on (HPI), by applying it to the counterexamples from §3.

5.1 Switching

It is condition (DIF) that allows (PART) to deliver the correct verdict in switching cases. Reconsider example **Switch** (§6, p. 16). In this example, every theory we considered delivered the verdict that $S = 1$ is an actual cause of $I = 1$. Not so for (PART). If we hold off-path variable L1 fixed at the actual value $L1 = 0$, setting $S = 0$ would change I to value $I = 0$. But it is also the case that in the state of the model produced by setting $S = 0$, if we hold off-path variable L2 fixed at the new actual value $L2 = 0$, setting $S = 1$ would change I to value $I = 0$. So (DIF) is violated and $S = 1$ is not an actual cause of $I = 1$. Similar reasoning shows that in example **Shock** (§7, p. 17), $A = 1$ is not an actual cause of $C = 1$ according to (PART).

McDermott (1995, "The Push", pp. 524*ff*, p. 538) and Hall (2000, "The Kiss", p. 209) have given examples which seem to show that (DIF) is too strong. Here I will only discuss McDermott's example, since everything I say also applies to Hall's. Here is how McDermott introduces "The Push":

> Suppose I push Jones in front of a truck, which hits him and kills him; if I had not done so, he would have been hit and killed by a bus. Common sense says that my push was a cause of his death. But the death would have occurred without my push [...]

There is a superficial resemblance between this case and the switching cases in §3.1. In all of these cases, there are two possibilities, both of which would have led to the same

[31] Principles governing model appropriateness that appeal to relevant possibilities have been proposed by Hitchcock (2001, p. 298), Woodward (2003, §2.8; 2008, §6) and Halpern and Pearl (2005, §5). The appeal to relevant possibilities in the analysis of causation is also made by Collins (2000), Lewis (2000), McDermott (2002) and Schaffer (2005).

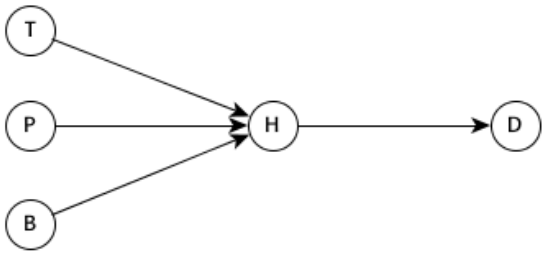
[32] See Appendix: (PART) Applied (p. 37).

Push
I push ($P = 1$) Jones in front of a truck ($T = 1$), which hits ($H = 1$) and kills him ($D = 1$); if I had not done so ($P = 0$), a bus ($B = 1$) would have hit ($H = 1$) and killed him.

Version (A): There is no other action available.
Version (B): I can push ($P = 2$) Jones to safety ($H = 0$).

$$P := 1, T := 1, B := 1 \qquad \text{(EX)}$$

$$H := (P = 1 \wedge T = 1) \vee (P = 0 \wedge B = 1), D := H \qquad \text{(ED)}$$



Example 12: *Directed Graph for Push*

result. And so it seems at first glance that if (DIF) rules out causation in those cases, it must also rule out causation in this one. Not so. Notice that it was appropriate to model the earlier cases with a switch variable that only had two possible values, representing the fact that the two possibilities were appropriately treated as exhaustive. I claim that in McDermott's case, our intuition that the push was a cause of death is best explained by the fact that we do not treat the particular pushing that occurred, and the alternative non-pushing, as exhausting the relevant possibilities. Rather, we suppose there must have been some alternative relevant possibility for action, such that had that possibility been realised the death would not have occurred. If so, then according to *Relevance* the model that reflects the case should have three variable values for the push, and it will then not be excluded as a cause by (DIF). That is, I hypothesise that our intuitions are reflecting the model for version (**B**) rather than for version (**A**) in example **Push** (§12, p. 26).

This hypothesis predicts that our intuitions can be shifted as follows. First, suppose we add to the story that any possible action I could have taken would have led to the death. Then the case is appropriately modelled by **Push** (**A**), is relevantly like the cases in §3.1, and our intuitions will say the push was not a cause of death. Second, suppose instead that we add to the story that there were three possible actions available: to not push, to push to safety, and to push in front of the truck. Then the case is appropriately modelled by **Push** (**B**), is not relevantly like the cases in §3.1, and our intuitions will say that pushing in front of the truck was a cause of death. I claim that these predictions are correct.

Notice that in reasoning in this way, at no time was it relevant to consider what would have happened had the candidate cause not occurred. Both McDermott and Hall, in describing their cases, suppose that it is sufficient to fix the causal facts of the situation to specify what would have happened had the candidate cause not occurred. But according to the hypothesis above this is not the case, for we also need to know the full range of alternative possibilities it is appropriate to consider, regardless of which was most likely to occur. This is because in (PART), like all theories of actual causation in the causal modelling framework, the counterfactuals that must be true for causal claims to be true do not concern claims about which variable settings would have been most likely to obtain were they different, but rather concern the results of hypothetical interventions. In just this sense then, I claim that the cases of McDermott and Hall are underspecified. Either we spell them out so that the options they mention are exhaustive or only omit possibilities that would also have led to the same result, in which case the switch is not an actual cause; or we spell them out so that the mentioned options are not exhaustive and omit some possibility that would have led to a different

result, in which case the switch is an actual cause. Either way, (DIF) survives[33].

5.2 Trumping

It is condition (STRAND) that allows (PART) to deliver the correct verdict in trumping cases. Reconsider example **Command** (§8, p. 18). In this example, every theory we considered except (PRE) delivered the verdict that $S = 1$ is an actual cause of $C = 1$. Not so for (PART). The only minimal set for $C = 1$ is the singleton $\{M = 1\}$, so (STRAND) is violated for $S = 1$, which therefore is not an actual cause of $C = 1$ according to (PART)[34].

McDermott (2002) argues that $S = 1$ *is* an actual cause of $C = 1$ on grounds that *ceteris paribus* causation is intrinsic, and that here *ceteris* is *paribus*. He then argues for a theory that delivers exactly this result. I will first argue against the theory, and then argue *ad hominem* that McDermott himself has given us reason to think that here *ceteris* is not *paribus*.

McDermott's theory counts $S = 1$ an actual cause of $C = 1$ because it is part of a minimal sufficient condition in the following sense (p. 97):

> A minimal sufficient condition for E is a sufficient condition in which no conjunct could be replaced by a weaker condition on what happens at that point without losing sufficiency.

McDermott says that $S = 1$ is part of the minimal sufficient condition $(S = 1) \wedge (M \neq 2)$. Notice that this condition cannot be a minimal set in the sense employed by (PART), since a minimal set involves the specification of the actual values of the

[33]Hall (2000, p. 207) also discusses a case where there are only two relevant possibilities, but in which the two paths to the effect are significantly different. I agree with Hall that our intuitions are shifted in this case, but in my view it is because it is described in a way which suggests that the chances of the effect are significantly different along the two possible routes. Cases like this are best modelled by probabilistic equations, and hence fall outside the scope of this paper. A referee suggests that an alternative explanation of our judgements in cases like those discussed in this section might be provided in terms of an interaction between moral judgement and causal judgement, of the sort proposed by Hitchcock and Knobe (2009). I hope to explore the relationship between causal and normative judgement on another occasion.

[34]Halpern and Pearl (2005, p. 874) note this asymmetry between the two candidate causes, but do not employ it in their treatment of trumping. The treatment of trumping I have given here will be available whenever it is appropriate to model the situation such that the variables representing the trumping and trumped events can conflict, with the trumping event determining the outcome (Field 2003, p. 452, fn. 27 and Halpern and Hitchcock 2010, §4.2 also note the importance of this fact). Schaffer has suggested (personal communication to Woodward) that even in cases which are not appropriately modelled in this way, trumping is a species of preemption. I disagree (see Woodward 2003, pp. 81–82, esp. fn. 45).

relevant variables, and $(S = 1) \wedge (M \neq 2)$ fails to specify an actual value for M. Of course, the question to ask at this point is why we should not permit logical operations on variable values in the specification of sufficient conditions. Indeed, it might be thought that since in general we should permit disjunctive causes (Sartorio 2006), we have no good reason to rule out logical operations in the specification of sufficient conditions. I agree, and leave open the possibility that (PART) might be generalised to permit logical operations on variables in the specification of causes and sufficient conditions. However, in this case the logical weakening permitted by McDermott's definition is guilty of generating the wrong result. For note that condition $(S = 1) \wedge (M \neq 2)$ is satisfied in virtue of condition $(S = 1) \wedge (M = 1)$ being satisfied, and in this latter condition $(S = 1)$ is redundant. We should not call a condition minimally sufficient if there exists a condition in virtue of which it could obtain that is not itself minimally sufficient.

McDermott endorses the verdict of his theory because he thinks we should believe that causation is intrinsic except in cases of double prevention (p. 89). In an otherwise identical variant of **Command** where there is no major present, the sergeant's order would be an actual cause of the corporal's charge. Since trumping cases are not cases of double prevention, McDermott suggests, we should therefore say that the order is also an actual cause in **Command**[35]. But as McDermott himself points out (p. 98), being a minimal set is not an intrinsic property of a set of variable values. McDermott's definition of sufficiency is relativised to what he calls "relevant variables". In our framework, the relevant variables consist of all variables in an appropriate model. For McDermott, as for us, sufficiency is defined in terms of whether the effect would still have occurred were relevant variables different. So whether a set of variable values is sufficient for another depends on which variables are relevant. Add the claim that the same sequence of event types could occur in contexts that differ concerning the relevant variables, and we have the result that being a sufficient condition is not an intrinsic property of a set of variable values. Here is how this applies to **Command**. In a situation where the major is not present, it is not appropriate to include a variable representing the possibility of the major giving an order. In such a context, the sergeant's order will be sufficient for the corporal's behaviour. But in a situation where the major is present, it is appropriate to include a variable representing the possibility of the major giving an order. In such a context, the sergeant's order will not be sufficient for the corporal's behaviour. So on McDermott's own account, as on ours, trumping cases provide contexts where causation is not intrinsic.

[35]Similar arguments are sometimes made in terms of the idea that the "causal process" connecting the sergeant and corporal is "complete" (Hitchcock 2011, §4).

5.3 Combination Lamp

It is condition (STRAND) that allows (PART) to deliver the correct verdict in **Loader** and **Combination Lamp**. In both of these examples, $A = 1$ is not an actual cause for the simple reason that it is not on a strand.

This result is the majority verdict among those I have informally polled. However, Joe Halpern and Chris Hitchcock (personal communication) have suggested that the actual causal facts in **Combination Lamp** are underdetermined by the description of the example. In particular, they have suggested that there exists a possible way of implementing the example in which, they claim, $A = 1$ is a cause of $L = 1$. Consider **Fancy Lamp** (§13, p. 31), suggested to me by <blinded>. In this example, they claim, we should judge that $A = 1$ is a cause of $L = 1$. I disagree. If **Combination Lamp** is implemented with the structure of **Fancy Lamp**, then $A = 1$ is on a strand for $L = 1$. And an intervention setting $A = 0$ satisfies all of the conditions in (PART)—except for (DIF). For in the state of the model produced by setting $A = 0$, an intervention setting $A = 1$ also satisfies those conditions. So according to (PART), it does not matter whether or not **Combination Lamp** is implemented in the manner of **Fancy Lamp**. Either way, $A = 1$ is not a cause of $L = 1$. In my view this is the correct result. In **Fancy Lamp** the position of switch A makes a difference to *how* L is brought about but does not make a difference to *whether* L is brought about.

6 Conclusion

I have argued that (PART) is the correct partial theory of actual causation, in the sense that it provides exactly those conditions required to rule out all non-causes that can be discerned at the level of counterfactual structure. A complete theory of actual causation requires substantially more work. First, it requires an account of the point of having a concept with just these contours. It is striking that we have a concept that cuts as finely as the differences between the theories of actual causation considered in this paper. Why should we have a concept that picks out *this* particular aspect of counterfactual structure? Second, it requires a comprehensive theory of the principles governing model appropriateness. Third, it requires the addition of conditions that can deliver the correct verdict for the class of non-structural counterexamples. If I am right that (PART) suffices for the rest, then there is a nice irony in the fact that the most plausible counterfactual theory of causation turns out to draw so heavily from the resources of the regularity theories it was initially motivated by rejecting.

Fancy Lamp

A lamp (L) is controlled by three switches (A, B and C), each of which has three possible positions $(-1, 0, 1)$. The switches are connected to detectors (N_{-1}, N_0, N_1), each of which is activated *iff* no switch is in position $(-1, 0, 1)$ respectively. The lamp switches on *iff* some detector is activated. In fact, the switches are in positions $A = 1$, $B = -1$ and $C = -1$, detector N_0 is activated, and $L = 1$.

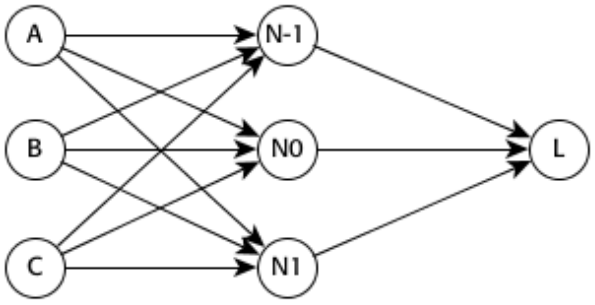
$$A := 1, B := -1, C := -1 \tag{EX}$$

$$N_{-1} := \neg(A = -1 \vee B = -1 \vee C = -1) \tag{ED}$$

$$N_0 := \neg(A = 0 \vee B = 0 \vee C = 0)$$

$$N_1 := \neg(A = 1 \vee B = 1 \vee C = 1)$$

$$L := N_{-1} \vee N_0 \vee N_1$$



Example 13: *Directed Graph for Fancy Lamp*

References

Richard A. Baldwin and Eric Neufeld. 2004. "The Structural Model Interpretation of the NESS Test", in *Advances in Artificial Intelligence: 17th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2004, London, Ontario, Canada, May 17-19, 2004. Proceedings*, edited by Ahmed Y. Tawfik and Scott D. Goodwin. Vol. 3060, Lecture Notes in Computer Science, Springer, Berlin, pp. 292–307. URL: http://dx.doi.org/10.1007/b97823.

Gunnar Björnsson. 2007. "How Effects Depend on their Causes, Why Causal Transitivity Fails, and Why we Care About Causation", in *Philosophical Studies*, Vol. 133, No. 3, Apr. 2007, pp. 349–390. URL: http://dx.doi.org/10.1007/s11098-005-4539-8.

John Collins. 2000. "Preemptive Prevention", in *The Journal of Philosophy*, Vol. 97, No. 4, Special Issue: Causation, Apr. 2000, pp. 223–234. Reprinted in Collins, Hall, and Paul (2004, pp. 107–118).

John Collins, Ned Hall, and L. A. Paul. 2004. *Causation and Counterfactuals,* edited by John Collins, Ned Hall, and L. A. Paul. Cambridge MA: MIT Press.

Hartry Field. 2003. "Causation in a Physical World", in *The Oxford Handbook of Metaphysics*, edited by Michael J. Loux and Dean W. Zimmerman, Oxford University Press, Oxford, pp. 435–460. URL: http://dx.doi.org/10.1093/oxfordhb/9780199284221.003.0015.

Ned Hall. 2000. "Causation and the Price of Transitivity", in *The Journal of Philosophy*, Vol. 97, No. 4, Special Issue: Causation 2000, pp. 198–222. Reprinted in Collins, Hall, and Paul (2004, pp. 181–204).

——— . 2004a. "The Intrinsic Character of Causation", in *Oxford Studies in Metaphysics*, edited by Dean W. Zimmerman. Vol. 1, Oxford University Press, Oxford, pp. 255–299.

——— . 2004b. "Two Concepts of Causation", in *Counterfactuals and Causation*, edited by John Collins, Ned Hall, and L. A. Paul, MIT Press, Cambridge MA, pp. 225–276.

——— . 2007. "Structural Equations and Causation", in *Philosophical Studies*, Vol. 132, No. 1, Jan. 2007, pp. 109–136. URL: http://dx.doi.org/10.1007/s11098-006-9057-9.

Joseph Y. Halpern. 2008. "Defaults and Normality in Causal Structures", in *Eleventh International Conference on Principles of Knowledge Representation and Reasoning (KR2008)*, edited by Jérôme Lang and Gerhard Brewka, AAAI Press, Menlo Park CA, pp. 198–208. URL: http://arxiv.org/abs/0806.2140.

Joseph Y. Halpern and Christopher Hitchcock. 2010. "Actual Causation and the Art of Modeling", in *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, edited by Rina Dechter, Hector Geffner, and Joseph Y. Halpern, College Publications, London, pp. 383–406.

Joseph Y. Halpern and Judea Pearl. 2001. "Causes and Explanations: A Structural-Model Approach. Part I: Causes", in *Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference*, edited by Jack Breese and Daphne Koller, Morgan Kaufmann, San Francisco, pp. 194–202. Revised version published as Halpern and Pearl (2005). URL: http://arxiv.org/abs/cs/0011012v2.

———. 2005. "Causes and Explanations: A Structural-Model Approach. Part I: Causes", in *British Journal for the Philosophy of Science*, Vol. 56, No. 4, Dec. 2005, pp. 843–887. Revised version of Halpern and Pearl (2001). URL: http://dx.doi.org/10.1093/bjps/axi147.

Daniel M. Hausman. 2005. "Causal Relata: Tokens, Types, or Variables?", in *Erkenntnis*, Vol. 63, No. 1, July 2005, pp. 33–54. URL: http://dx.doi.org/10.1007/s10670-005-0562-6.

Eric Hiddleston. 2005. "Causal Powers", in *The British Journal for the Philosophy of Science*, Vol. 56, No. 1, Mar. 2005, pp. 27 –59. URL: http://dx.doi.org/10.1093/phisci/axi102.

Christopher Hitchcock. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs", in *The Journal of Philosophy*, Vol. 98, No. 6, June 2001, pp. 273–299.

———. 2007. "Prevention, Preemption, and the Principle of Sufficient Reason", in *Philosophical Review*, Vol. 116, No. 4, Oct. 2007, pp. 495–532. URL: http://dx.doi.org/10.1215/00318108-2007-012.

———. 2011. "Trumping and Contrastive Causation", in *Synthese*, Vol. 181, No. 2, July 2011, pp. 227–240. URL: http://dx.doi.org/10.1007/s11229-010-9799-y.

Christopher Hitchcock and Joshua Knobe. 2009. "Cause and Norm", in *Journal of Philosophy*, Vol. 106, No. 11, Nov. 2009, pp. 587–612.

Mark Hopkins and Judea Pearl. 2003. "Clarifying the Use of Structural Models for Commonsense Causal Reasoning", in *Logical Formalization Of Commonsense Reasoning*, edited by Patrick Doherty, John McCarthy, and Mary-Anne Williams, Technical Report (American Association for Artificial Intelligence) SS-03-05, AAAI Press, Menlo Park, CA, pp. 83–89. URL: http://www.aaai.org/Papers/Symposia/Spring/2003/SS-03-05/SS03-05-011.pdf.

Jérôme Lang and Gerhard Brewka. 2008. *Eleventh International Conference on Principles of Knowledge Representation and Reasoning (KR2008),* edited by Jérôme Lang and Gerhard Brewka. Menlo Park CA: AAAI Press.

Marc Lange. 2006. *Philosophy of Science: An Anthology,* edited by Marc Lange. Malden MA: Blackwell.

David Lewis. 1973. "Causation", in *The Journal of Philosophy*, Vol. 70, No. 17, Oct. 1973, pp. 556–567. Reprinted in Lewis (1986, pp. 159–171) and Tooley (1999, pp. 178–189). URL: http://dx.doi.org/10.2307/2025310.

———. 1986. *Philosophical Papers,* Vol. II. Oxford: Oxford University Press. URL: http://dx.doi.org/10.1093/0195036468.001.0001.

———. 2000. "Causation as Influence", in *The Journal of Philosophy*, Vol. 97, No. 4, Special Issue: Causation 2000, pp. 182–197. Reprinted in Collins, Hall, and Paul (2004, pp. 75–106) and Lange (2006, pp. 466–487). URL: http://dx.doi.org/10.2307/2678389.

John L. Mackie. 1974. *The Cement of the Universe: A Study of Causation,* Oxford: Oxford University Press. URL: http://dx.doi.org/10.1093/0198246420.001.0001.

Tim Maudlin. 2004. "Causation, Counterfactuals, and the Third Factor", in *Counterfactuals and Causation*, edited by John Collins, Ned Hall, and L. A. Paul, MIT Press, Cambridge MA, pp. 419–443. Reprinted in Maudlin (2007, pp. 143–169). URL: http://dx.doi.org/10.1093/acprof:oso/9780199218219.003.0006.

———. 2007. *The Metaphysics Within Physics,* Oxford: Oxford University Press. URL: http://dx.doi.org/10.1093/acprof:oso/9780199218219.001.0001.

Michael McDermott. 1995. "Redundant Causation", in *British Journal for the Philosophy of Science*, Vol. 46, No. 4, pp. 523–544. URL: http://dx.doi.org/10.1093/bjps/46.4.52.

———. 2002. "Causation: Influence versus Sufficiency", in *The Journal of Philosophy*, Vol. 99, No. 2, Feb. 2002, pp. 84–101.

Peter Menzies. 2004. "Causal Models, Token Causation, and Processes", in *Philosophy of Science*, Vol. 71, No. 5, Dec. 2004, pp. 820–832. Expanded version available as a preprint at http://philsci-archive.pitt.edu/1039/. URL: http://dx.doi.org/10.1086/425057.

———. 2007. "Causation in Context", in *Causation, Physics and the Constitution of Reality: Russell's Republic Revisited*, edited by Huw Price and Richard Corry, Oxford University Press, Oxford, pp. 191–223.

James D. Park. 2003. "Causes and Explanations Revisited", in *IJCAI-03: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, edited by Georg Gottlob and Toby Walsh, Morgan Kaufmann, San Francisco, pp. 154–159.

Judea Pearl. 2000. *Causality,* Cambridge: Cambridge University Press.

———. 2009. *Causality,* 2nd edition. Cambridge: Cambridge University Press.

Carolina Sartorio. 2005. “Causes As Difference-Makers”, in *Philosophical Studies*, Vol. 123, No. 1-2, Mar. 2005, pp. 71–96. URL: http://dx.doi.org/10.1007/s11098-004-5217-y.

——— . 2006. “Disjunctive Causes”, in *The Journal of Philosophy*, Vol. 103, No. 10, Oct. 2006, pp. 521–538.

Jonathan Schaffer. 2000. “Trumping Preemption”, in *The Journal of Philosophy*, Vol. 97, No. 4, Special Issue: Causation 2000, pp. 165–181. Reprinted in Collins, Hall, and Paul (2004, pp. 59–74).

——— . 2005. “Contrastive Causation”, in *The Philosophical Review*, Vol. 114, No. 3, Jan. 2005, pp. 327–358.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction and Search,* 2nd edition. Cambridge MA: MIT Press. URL: http://cognet.mit.edu/library/books/view?isbn=0262194406.

Michael Strevens. 2007. “Mackie Remixed”, in *Causation and Explanation*, edited by Joseph Keim Campbell, Michael O’Rourke, and Harry S. Silverstein. Vol. 4, Topics in Contemporary Philosophy, MIT Press, Cambridge MA, pp. 93–118.

——— . 2008. *Depth: An Account of Scientific Explanation,* Cambridge MA: Harvard University Press.

Michael Tooley. 1999. *Laws of Nature, Causation, and Supervenience,* edited by Michael Tooley. Vol. 1. Metaphysics. New York: Garland.

Brad Weslake. ms. “Exclusion Excluded”. Under review. URL: http://philpapers.org/rec/wesee.

James Woodward. 2003. *Making Things Happen: A Theory of Causal Explanation,* New York: Oxford University Press. URL: http://dx.doi.org/10.1093/0195155270.001.0001.

——— . 2008. “Response to Strevens”, in *Philosophy and Phenomenological Research*, Vol. 77, No. 1, Aug. 2008, pp. 193–212. URL: http://dx.doi.org/10.1111/j.1933-1592.2008.00181.x.

Stephen Yablo. 2002. “De Facto Dependence”, in *The Journal of Philosophy*, Vol. 99, No. 3, Mar. 2002, pp. 130–148.

——— . 2004. “Advertisement for a Sketch of an Outline of a Prototheory of Causation”, in *Counterfactuals and Causation*, edited by John Collins, Ned Hall, and L. A. Paul, MIT Press, Cambridge MA, pp. 119–138. Reprinted in Yablo (2010, pp. 98–116).

——— . 2010. *Things: Papers on Objects, Events, and Properties,* Oxford: Oxford University Press. URL: http://dx.doi.org/10.1093/acprof:oso/9780199266487.001.0001.

Dean W. Zimmerman. 2004. *Oxford Studies in Metaphysics,* edited by Dean W. Zimmerman. Vol. 1. Oxford: Oxford University Press.

Appendix: (PART) Applied

	Possible Causes			Non-Causes	
Example	Variable	Path	(PART$_P$)	Variable	Violated
Backup (§1, p. 8)	$T = 1$	$T \to V$	$\{S = 0\}$	$S = 0$	(STRAND)
Window (§2, p. 9)	$B = 1$	$B \to W$	$\{S = 0\}$		
	$S = 1$	$S \to W$	$\{B = 0\}$		
Bottle (§3, p. 11)	$ST = 1$	$ST \to SH \to BS$	$\{BH = 0\}$	$BT = 0$	(STRAND)
	$SH = 1$	$SH \to BS$	$\{BH = 0\}$	$BH = 0$	(STRAND)
Vote (§4, p. 13)	$V_1 = 1$	$V_1 \to M \to P$	$\{V_2 = 0\}$		
	$V_2 = 1$	$V_2 \to M \to P$	$\{V_1 = 0\}$		
	$M = 2$	$M \to P$	$\{\}$		
Loader (§5, p. 14)	$C = 1$	$C \to D$	$\{A = 1,$	$A = 1$	(STRAND)
			$B = 0\}$	$B = 0$	(STRAND)
Switch (§6, p. 16)	$L2 = 1$	$L2 \to I$	$\{L1 = 0\}$	$L1 = 0$	(STRAND)
				$S = 1$	(DIF)
Shock (§7, p. 17)	$B = 1$	$B \to C$	$\{A = 1\}$	$A = 1$	(DIF)
Command (§8, p. 18)	$M = 1$	$M \to C$	$\{S = 1\}$	$S = 1$	(STRAND)
Lamp (§9, p. 19)	$B = -1$	$B \to L$	$\{A = 1,$	$A = 1$	(STRAND)
			$C = -1\}$		
	$C = -1$	$C \to L$	$\{A = 1,$		
			$B = -1\}$		
Antidote* (§10, p. 20)	$A = 0$	$A \to D$	$\{B = 0\}$		
	$B = 1$	$B \to D$	$\{A = 1\}$		
Poisoning* (§11, p. 22)	$A = 1$	$A \to D$	$\{B = 1\}$	$B = 1$	(STRAND)
Push (§12, p. 26)					
Version (A)	$T = 1$	$T \to H \to D$	$\{P = 1\}$	$B = 1$	(STRAND)
	$H = 1$	$H \to D$	$\{\}$	$P = 1$	(DIF)
Version (B)	$T = 1$	$T \to H \to D$	$\{P = 1\}$	$B = 1$	(STRAND)
	$P = 1$	$P \to H \to D$	$\{\}$		
	$H = 1$	$H \to D$	$\{\}$		
Fancy (§13, p. 31)	$B = -1$	$B \to N_0 \to L$	$\{A = 1,$	$A = 1$	(DIF)
			$C = -1,$	$N_1 = 0$	(STRAND)
			$N_{-1} = 0,$	$N_{-1} = 0$	(STRAND)
			$N_1 = 0\}$		
	$C = -1$	$C \to N_0 \to L$	$\{A = 1,$		
			$B = -1\}$		
			$N_{-1} = 0,$		
			$N_1 = 0\}$		

Notes
Non-structural counterexamples are marked with (*).