# Causal Loops and the Reduction of Causation

Jens Jäger

*11,650w.*

Draft: 12 September 2024

**Abstract**

One of Hume's definitions of causation reduces it to counterfactual dependence: according to Hume, a cause is nothing but "an object followed by another ... where, if the first object had not been, the second would never have existed". Modern attempts to reduce causation to counterfactuals have long abandoned one half of Hume's proposal: it's nearly universally accepted that counterfactual dependence isn't *necessary* for causation—sometimes $A$ is a cause of $B$ even though $B$ would still be if $A$ wasn't. But the other half has received much less criticism: it is still often maintained that counterfactual dependence is *sufficient* for causation—that $A$ causes $B$ provided that $B$ would not be if $A$ wasn't. This essay has two goals. First, I argue that even weak versions of this sufficiency claim are false, based on a case involving causal loops. As a result, any counterfactualist reduction of causation entailing the sufficiency claim should be rejected—this includes the seminal accounts of Lewis and of Halpern and Pearl. Now, as it happens, some structural equation accounts *don't* entail the sufficiency claim. However, I argue second, those accounts still don't satisfactorily handle causal loop scenarios, because the reductive accounts of structural equations they rely on misfire in the presence of causal loops. I'll finish with some remarks about the way forward.

# Contents



# Introduction

Let an "event proposition" be any proposition apt to figure as a cause or effect. For concreteness, I'll let an event proposition be any proposition of the form $Q(R)$, where $R$ is a spacetime region and $Q$ a possible intrinsic property of $R$—but other reasonable choices will also work.[1] Say that two event propositions are *disjoint* iff they are entirely about disjoint regions of spacetime. Any two disjoint event propositions describe "wholly distinct" events.

Some counterfactual reductions of causation entail that counterfactual dependence between disjoint event propositions is *sufficient* for causation. If I weren't typing away at my keyboard, then no letters would appear on my screen. So, one concludes, typing away at my keyboard *causes* the letters to appear in on my computer screen. Generalizing:[2]

> **Sufficiency:** Necessarily, if $P$ and $Q$ are true disjoint event propositions such that it wouldn't be the case that $Q$ if it wasn't the case that $P$, then $P$ is a cause of $Q$.

Counterfactual conditionals are notoriously context-sensitive, and **Sufficiency** has initial plausibility only in some of those contexts. Lewis (1979, p. 458) famously distinguishes

[1] In particular, you might impose various "niceness" requirements, such as the properties' being non-disjunctive (Lewis, 1986; Jäger, 2021). More on this later.

[2] If you reject the idea that propositions are causal relata, it's easy to translate the below into the language of events: necessarily, for any disjoint $R$ and $S$, if $P = U(R)$ and $Q = V(S)$ and $U$ and $V$ are actual intrinsic—and otherwise sufficiently "nice" (non-disjunctive, not too fragile, etc)—properties of $R$ and $S$, respectively, and it wouldn't be the case that $Q$ if it wasn't the case that $P$, then $R$'s having $U$ is a cause of $S$'s having $V$. For convenience, though, I'll stick with the propositional ideology throughout.

between "standard" and "backtracking" contexts. Suppose that earlier today you returned a book to your friend Susy, as you promised you would do. Given that Susy is known to take promises seriously, the following conditional seems true:

(1) If I hadn't returned the book today, Susy would be disappointed in me.

From (1), **Sufficiency** would conclude that my returning the book was a cause of my staying in Susy's good graces. So far so good.

But now consider that you're extremely reliable and honest, known to never break a promise. With this in mind, you could have reasoned as follows:

(2) If I hadn't returned the book today, that would have been because Susy and I agreed on a later return date to begin with. So Susy wouldn't have been disappointed in me.

From (2), **Sufficiency** would conclude that your returning the book today is a cause of your agreeing on today as the return date. But that's absurd: you have no retrocausal powers.

Conditional (1) and conditional (2) exemplify, respectively, the "standard" and the "backtracking" reading of the counterfactual conditional. For just the aforementioned reasons, Lewis requires that the counterfactual conditional in **Sufficiency** be given its standard reading. We, too, shall heed this requirement.

The paper has two goals. First, I present a new argument against **Sufficiency**, involving a counterexample with causal loops. After laying out the argument and defending it, I show how it defeats several popular counterfactual reductions of causation, including Lewis (1973a), Halpern and Pearl (2005), and Hall (2007). *Some* counterfactual reductions, however, survive: in particular, some recent reductions of causation to structural equations don't entail **Sufficiency** and therefore avoid the particular counterexample. However, for a *complete* reduction of causation, any such reduction must be combined with a reduction of the structural equations. But, I argue secondly, no extant reduction of structural equations yields a winning combination: they all misfire in mild variations of the causal loop scenario. I'll finish with some remarks about the way forward.

# 1 Previous Arguments Against Sufficiency

Some take **Sufficiency** to be the cornerstone of our modern inquiry into causation. Here are Beckers and Vennekens (2017, p. 2, my emphasis):

> "The currently most prominent approaches to defining actual causation are those within the counterfactual dependence tradition, which started with Lewis (1973a). *All of these approaches take as their starting point the assumption that counterfactual dependence is sufficient for causation*, but not necessary (Hitchcock (2001); Woodward

(2003); Hall (2004; 2007); Halpern and Pearl (2005); Halpern (2016); Weslake (2015) ...)."

(We'll later see that several of these attributions are incorrect.)

But it hasn't been all praise for **Sufficiency**—there are a few ways to resist it. To start, one might think that omissions aren't causes, as Beebee (2004) argues. If Beebee is right, then since events can still counterfactually depend on omissions—like a plant's death on Flora's failure to water it—counterfactual dependence isn't sufficient for causation. But Beebee's arguments can be resisted in various ways. You might simply bite the bullet: even far-flung propositions, like Julius Caesar's failure to water Flora's plant, are a cause of the plant's death. You might try to blunt the bullet's impact in various ways, e.g. by partially reducing omissions to "positive events", i.e. *commissions* (Bernstein, 2014), or by explaining the appearance of non-causation as mere infelicity (Schaffer, 2005). As a last resort, you could also retreat to a weakening of **Sufficiency**, additionally requiring that $P$ and $Q$ be *positive* event propositions.

Proportional causation poses another *prima facie* challenge to **Sufficiency**. I greet my neighbor loudly, and she startles. My *greeting loudly* causes the startle, but my greeting *simpliciter* doesn't—my neighbor isn't *that* jumpy. Yet, if I hadn't greeted her *simpliciter*, my neighbor wouldn't have startled. So counterfactual dependence isn't sufficient for causation. To resist this line of reasoning, one could appeal to pragmatics, perhaps pointing out that mentioning that $A$ causes $B$ tends to carry the implicature that $A$ is a *maximally specific* cause of $B$. Or one could try to separate causation from explanation and shift the burden of proportionality over to the explanatory side. In any case, as a last resort, you could retreat to a weakening of **Sufficiency**, additionally requiring that $P$ and $Q$ to be *proportional* event propositions.

A third challenge emerges from Lewis's own miracles-based semantics for counterfactuals. The semantics doesn't in fact perfectly heed the distinction between the "standard" and "backtracking" resolution itself. Lewis grants that, even in standard contexts, the closest possible antecedent world may include a *ramping period*. On his account, counterfactual antecedents are preferentially brought about by small miracles—Lewis's closeness ranking most harshly penalizes large miracles—but small miracles need time to snowball into big change. So, where antecedents dictate big differences to actuality, there's typically a significant delay between miracle and antecedent event, during which the counterfactual world differs from actuality.

For example: you throw a ball at me. I notice just in time and catch it. If I hadn't caught the ball, surely that wouldn't be because at the moment of impact a miracle instantly twisted my arm, making me drop the ball. Instead, it would have been because I noticed slightly later, or you threw the ball a littler harder, or a gust of wind deflected the ball—some

non-instantaneous macroscopic change. Such a change could plausibly be brought about by a small miracle, e.g. by slightly altering neural firing patterns, or by minuscule past meteorological changes, eventually leading to the additional gust of wind. The period from whenever the miracle occurs to my failing to catch the ball is the counterfactual "ramping period". The need for ramping periods on the standard resolution raises, again, the specter of retrocausation: we certainly want to avoid saying that my failure to catch the ball *causes* my earlier neurons' firing, or *causes* the earlier atmospherical state.

Lewis (1979) offers a response. According to **Sufficiency**, $P$ causes $Q$ if $\neg P \mathrel{\Box\!\!\rightarrow} \neg Q$. On Lewis's semantics, this requires that $\neg Q$ in *all* closest worlds where $\neg P$. But the negation of an event proposition—like my catching the ball—typically leaves much undetermined. Lewis's hope is that this includes the ramping period's content too:

> "[W]e should sacrifice the independence of the immediate past to provide an orderly transition from actual past to counterfactual present and future. *That is not to say, how- ever, that the immediate past depends on the present in any very definite way*. There may be a variety of ways the transition might go, hence there may be no true counterfactuals that say in any detail how the immediate past would be if [some event proposition hadn't occurred]."[3] (Lewis, 1979, p.463, my emphasis)

This response works for our example. As we saw, there are all sorts of reasons I might not have caught the ball—heightened alertness, a harder throw, an altered wind pattern, etc. Each reason results in a different ramping period. A disjunction of such a diversity of ramping periods would reasonably count as too disjunctive to be a single event proposition. One might reasonably hope, therefore, that $\neg P \mathrel{\Box\!\!\rightarrow} \neg Q$ is false whenever $Q$ precedes $P$ and $P$ and $Q$ are both event propositions.

Vihvelin (1995) identifies two kinds of threats to this response. The first arises when the event proposition $P$ is an *omission*. But this is no threat to the weakening of **Sufficiency** to positive event propositions. The second threat stems from $Q$s that entail the entire immediate past of $P$. Let $P$ be *I catch the ball at t*, for some time $t$. Let $Q$ describe the actual entire state of the world during some open time interval ending in $t$. Given that the laws are deterministic, necessarily $Q$ is false if $P$ is false, and so $\neg P \mathrel{\Box\!\!\rightarrow} \neg Q$. But we certainly *don't* want to say that my catching the ball causes the world to be as it actually is in the run-up to my catching the ball.

One possible response to this follows Lewis (1986) in banning overly "fragile" events. In our framework, those are propositions attributing overly complex properties to space-

[3]The original quote ends with "...if the past were different". This hope is stronger and, I think, unjustified. For example, it entails that "there may be no true counterfactuals that say in any detail how the immediate past would be" if some *omission* hadn't occurred. But the non-occurrence of an omission is a positive event, so we'd generally expect it to fix the past in a rather definite way. I suspect my substitution better captures what Lewis actually had in mind.

time regions. Lewis's justification for the ban is that our standard way of denoting event propositions, using "standard nominalizations", isn't nearly detailed enough to select these propositions. Inspired by Lewis's move, one might retreat to a further weakening of **Sufficiency**, additionally requiring that $P$ and $Q$ not be *overly detailed* event propositions.[4]

Now, a *fourth* challenge emerges from *rejecting* counterfactual miracles in favor of a view on which counterfactual worlds have different pasts all the way back. Versions of this view have been endorsed (among others) by Bennett (1984), Loewer (2007), Albert (2015), and Dorr (2016). The view holds that, assuming I blink only once, then if I had blinked twice instead, the world would have been microscopically different arbitrarily far back in time (Dorr, 2016). But surely distant retrocausation isn't as easy as blinking once. So **Sufficiency** is wrong. The **Sufficiency** advocate could resist this by sticking with Lewis's miracles-based semantics (with the aforementioned caveats). As a last resort, she can also retreat to a weakening of **Sufficiency**, which requires $P$ and $Q$ to additionally be *macroscopic* event propositions.

**Sufficiency** emerges significantly weakened, but it still makes interesting and substantive predictions: the canonical examples of causation, which also motivate counterfactual reductions, tend to involve positive, proportional, macroscopic, and not overly detailed events—stone throws, falling boulders, hurricanes, poisonings, and the like. So, the proponent of **Sufficiency** might still think of herself as occupying a true and substantive position. Unfortunately, as I'll argue, that appearance is illusory: even weakenings of **Sufficiency** are false.[5]

[4] There is a different response available too: one could retreat to a variation of **Sufficiency** where the relevant counterfactual circumstances don't require ramping periods. Glynn's (2013) account is an example of this (itself a reductive variant of Woodward (2003)). It's a combination of two ideas. First, stipulate a technical usage of "$\boxrightarrow{}$" according to which, when $A = Q(t)$ for a single time $t$, "$A \boxrightarrow{} B$" is evaluated using only *maximally-late* miracles; that is, $A \boxrightarrow{} B$ is true iff all closest $A$-worlds are $B$-worlds in which no miracles occur outside of $t$, and in which everything prior to $t$ is as it actually is, and everything after $t$ evolves according to the actual laws of nature. Second, say that $A$ *causes* $B$ if there is some truth $T$ solely about $t$ such that $\neg A$ and $T$ are metaphysically compossible and $\neg A \wedge T \boxrightarrow{} \neg B$, where $\boxrightarrow{}$ is evaluated in the technical way above. The role of $T$ is to suppress any unwanted effects which the $\neg A$-realizing miracle might bring about—consequences which affect $B$ via causal routes bypassing $\neg A$. Generally, therefore, $T$ will be secured by a highly complex miracle. For example, my bus is stuck in traffic, and my being on the bus right now ($A$) causes my being late to the meeting ($B$). Moreover, if I wasn't on the bus right now, I'd be on my bike ($\neg A$), speeding through grid-locked traffic and arriving on time. On the technical, "maximally-late" resolution, if I was on my bike right now, this would be because a miracle now teleported me there. But normally such a miracle would leave me extremely startled—so much so that (let's suppose) I'd crash, thus not arriving on time after all. Still, we want to say that my being on the bus ($A$) causes my being late ($B$). Glynn's account achieves this, because some $T$ entail that I'm calm and not disoriented, such that $\neg A \wedge T$ ensures that I'd arrive on time despite the sudden teleport. The account thereby secures the desired causal relation (that my being on the bus causes my being late), while avoiding any counterfactual ramping period. Now, importantly for our purposes, Glynn's account still falls to my counterexample, as I'll explain in fn. 18.

[5] Theories which posit global nomic constraints might also seem problematic for **Sufficiency**. Consider classical electrodynamics. Gauss's law dictates that the flux of the electric field through any (spatial) volume's boundary equals the amount of charge enclosed in the volume. This raises the worry that the following counterfactual is true:

My argument considers the oft-neglected case of *causal loops*: situations where a causal influence travels back in time and interacts with its source. Formally, they are chains of event propositions $(X_1, ..., X_n)$ such that, for all $i = 1, ..., n-1$, $X_i$ causes $X_{i+1}$, and $X_n$ causes $X_1$. Counterfactual accounts of causation often bracket the case of causal loops (e.g. Hitchcock (2001), Hall (2007), and Weslake (2015); two exceptions are Halpern and Pearl (2005) and Halpern (2016)). This neglect is in *prima facie* tension with the goal of providing an *analysis* of causation. The nature of causation is no contingent matter: an analysis should hold in all possible worlds, including ones with causal loops. (Indeed, even theories that don't aim for a reductive analysis of causation, such as Woodward (2003), would all else equal be better off with intensional adequacy.) Now of course, this assumes that causal loops are metaphysically possible; I'll soon argue that they are.

One might think that the causal loop case is spoils for the victor. Our intuitions about causal loops are murky, the thought goes, and so we should just accept the predictions of whatever account yields the best results in the acyclic case, where our intuitions are solid. But often enough our intuitions about causation and counterfactuals on causal loops are solid, too. To illustrate: let there be a single particle, with mass $m$, moving on a two-dimensional temporally oriented spacetime, rolled up, in the time-like direction, into a cylinder.[6]

---

> **(C)** If the amount of charge had been different over here, then the electric field far away would have been different.

If (C) is true, then **Sufficiency** would imply that Maxwellian electrodynamics involves widespread action at a distance—an unpalatable implication.

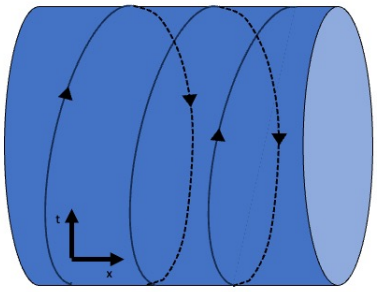
The **Sufficiency** proponent can reply that (C)'s antecedent is grossly underspecified. The truth of the consequent depends sensitively on the additional charges' origin. If the charges were transported over from somewhere else, far away field values would remain unchanged. *Only if* it the charges were somehow created *ex nihilo*, and without the simultaneous creation of nearby counterbalancing opposite charges, would the far-away electric field have been different.

But charge creation *ex nihilo*, without the creation of counterbalancing charges, is nomically prohibited by classical electrodynamics (and any of its successor theories). Thus *if* we favor counterfactual scenarios which preserve the actual laws (e.g. like Dorr (2016)), the world in which charges are merely moved, or created with counterbalancing charges, should be closer to actuality, and (C) therefore false.

Now, not everyone thinks that laws are counterfactually robust: Lewis (1973; 1979) famously proposes a miracle-based counterfactual semantics. But Lewis (1973b) also allows closeness ties between worlds (i.e., he rejects Conditional Excluded Middle (CEM)). This lightens the burden on the **Sufficiency** proponent. Now she need merely argue that worlds with unbalanced *ex nihilo* creation are never *strictly* closer to actuality than worlds in which charges are simply moved or created with counterbalancing charges. This seems much more feasible.

What about someone who likes Lewis's miracles view but also affirms CEM? Affirmation of CEM naturally goes along with a view that counterfactuals often have their truth value indeterminately. The **Sufficiency** proponent may then reasonably argue that it's never *determinately* the case that, if the local amount of charge had been different, then charge was created *ex nihilo* without counterbalancing charges. She can then retreat to a *further weakening* of **Sufficiency** which requires its counterfactual to have determinate truth value. This paper's argument also extends to this further weakening—our counterfactuals all have sufficiently specific antecedents.

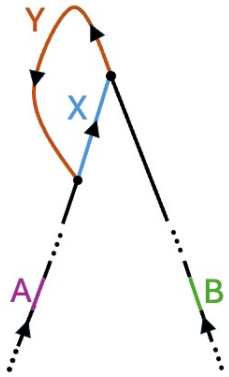
[6] Mathematically, we can represent this by a two-dimensional, oriented, connected, closed Lorentzian manifold.



Suppose the laws say that the particle moves uniformly, with (four-)velocity $\mathbf{u}$. It seems obvious—or at least as obvious as in the acyclic case—that the particle's being located at $\mathbf{x}$ *causes* its being located at $\mathbf{x}^+ := \mathbf{x} + \mathbf{u} \cdot \Delta\tau$ after (proper) time $\Delta\tau$. And it seems equally obvious that, if the particle had, at $\mathbf{x}$, a different velocity, $\mathbf{u}'$, it would instead be at $\mathbf{x}^+ := \mathbf{x} + \mathbf{u}' \cdot \Delta\tau$ after time $\Delta\tau$. Our intuitive judgments here seem clear.

## 2 Counterfactual Dependence on Loops

Now, unfortunately for the **Sufficiency** proponent, some of our intuitive judgments in causal loop cases conflict with her principle. Consider FORKS, a one-dimensional spacetime with the following topological structure:




The structure can be constructed by taking a one-dimensional oriented spacetime with an "inverse fork" and "identifying" the upper boundary (end point) of $Y$ with the lower boundary (starting point) of $X$. The figure indicates two additional intervals, $A$ and $B$, somewhere in the loop's distant past.

Suppose the spacetime is home to a scalar field, which takes 0 or 1 at each spacetime point. According to the dynamical laws, there are two kinds of spacetime points, *boring points* and *fork points*. Fork points are where two lines merge (indicated by black dots in the

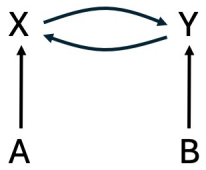
above figure).[7] Boring points are all the other points. In any interval composed wholly of boring points the field remains constant. Likewise for any interval starting in a fork point and otherwise composed of boring points. At fork points, meanwhile, the field's value is determined by the field's values on the incoming lines: if the values on each line agree, then the value at the fork point is 1; otherwise it is 0.

We can write these dynamics in FORKS more compactly, with the help of structural equations. Throughout, I'll use "structural equation" as a name for an expression of the form $\ulcorner X := f_X(Y_1, ..., Y_n)\urcorner$, such that $X, Y_1, ..., Y_n$ are variable names and $f_X$ represents any function from $n$-tuples of real numbers into the real numbers. I'll always interpret structural equations as implying that the values of $Y_1, ..., Y_n$ jointly "determine" the value of $X$—hence the use of the asymmetric symbol ":=". But different senses of "determine" will play a role in different places. Presently, the relevant sense of "determine" is *nomic determination*. (Later, we'll see that others want to interpret ":=" as something like *counterfactual determination*.)

For each region indicated above we introduce a binary variable, whose value represent the fields value taken at that region. For convenience, I'll reuse spacetime region names for variable names. The relations of nomic determination are captured by the following two structural equations. (For any $V$ and $W$, $V \leftrightarrow W$ is short for $W \cdot V + \overline{W} \cdot \overline{V}$, and $\overline{V}$ is short for $(1 - V)$.)

$$\begin{aligned} X &:= A \leftrightarrow Y \\ Y &:= B \leftrightarrow X \end{aligned} \tag{1}$$

A graphical representation helps to bring out the underlying symmetry. The graph's nodes are identified with the respective regions, and an arrow from one region to another signifies that the (variable representing the) former region appears on the right-hand side of the equation of the (variable representing the) latter:




Jointly the two equations in 1 have exactly four distinct solutions:[8]

[7] We can characterize them topologically as follows: a fork point is any point $p$ such that there are two open lines containing $p$ which don't share an open sub-line containing $p$.

[8] For a set of structural equations, a *solution* is an assignment of values to every variable which appears on the right-hand side of some equation. Given a solution $S$, the structural equations generate an assignment $S'$ of values to each variable appearing on the left-hand side of some equation. A solution is *consistent* iff, for any variable $V$ in the intersection of the domains of $S$ and $S'$, $S(V) = S'(V)$.

| $A$ | $B$ | $X$ | $Y$ |
|---|---|---|---|
| 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 |
| 1 | 1 | 1 | 1 |
| 1 | 1 | 0 | 0 |

Note that $A$ and $B$ are *perfectly correlated*—given the spacetime structure and the given dynamics, it's impossible for them to have different values.

Now suppose that, actually, $A = B = X = Y = 1$. I claim that, in this case, *if $A$ had been* 0, *then $B$ would have been* 0:

$$A = 0 \mathrel{\Box\!\!\rightarrow} B = 0,$$

where $\Box\!\!\rightarrow$ is the counterfactual conditional. **Sufficiency** then implies that $A = 1$ is a cause of $B = 1$. But, I say, $A = 1$ isn't a cause of $B = 1$. So Sufficiency is false.

As a valid argument (where the modalities are metaphysical):

1. If **Sufficiency** is necessarily true, FORKS is possible, $A = 1$ and $B = 1$ are disjoint event propositions in FORKS, and if $A = 0 \mathrel{\Box\!\!\rightarrow} B = 0$ in FORKS, then $A = 1$ causes $B = 1$ in FORKS.

2. $A = 1$ and $B = 1$ are disjoint event propositions in FORKS.

3. **Possibility:** FORKS is possible.

4. **Dependence:** $A = 0 \mathrel{\Box\!\!\rightarrow} B = 0$ in FORKS.

5. **Non-Causation:** $A = 1$ does *not* cause $B = 1$ in FORKS.

(C) **Sufficiency** is not necessarily true.

Premise 1 follows from the fact that a necessary truth is true in all possible worlds. Premise 2 follows because assignments of sufficiently "nice" intrinsic properties to regions are event propositions. Views will vary on what exactly niceness requires. Nice properties may be required to be "non-disjunctive" (Jäger, 2021); and with a view to last section's discussion, we may also prohibit them from being overly detailed (Lewis, 1986). To appease opponents of omissions, we require them to be positive, too. $A = 1$ and $B = 1$ tick all three of these boxes.

What about the other two weakenings of **Sufficiency**, to proportional and macroscopic event propositions? $A = 1$ and $B = 1$ are clearly proportional to each other, attributing as they do identical field values to equal-dimensional intervals. Unfortunately, they aren't "macroscopic" in the relevant sense: views like Dorr (2016) would insist that, if it was the case

that $A = 0$, then it would have to have been the case that the field on the $A$-branch is zero all the way back.

*Fortunately*, it's not hard to come up with a macroscopic analogue of our toy example. Here's one: replace the field by bowling balls, of which there are two types: green ones (1) and red ones (0). The dynamics dictate that, whenever two bowling balls collide, if their colors match, they fuse into a green bowling ball; otherwise they fuse into a red bowling ball. One day, Alice and Bob each hurl bowling balls toward a wormhole entrance. Alice's ball is off-course; luckily, a wormhole exit opens up and ejects a bowling ball, which collides with Alice's in just the right way that the resulting fusion product is back on course to hit the wormhole entrance. Shortly after, the fusion product reaches the wormhole entrance, simultaneously with Bob's bowling ball. The two balls collide and their fusion product enters the wormhole, being ejected a short while earlier, on course to collide with Alice's original bowling ball. We may suppose that, like in FORKS, the relevant nomic structure follows the equations 1, with $A$ and $B$ tracking the color of Alice's and of Bob's original bowling ball, respectively, $X$ tracking the color of the first fusion product, and $Y$ the color of the second fusion product. The first fusion product is green ($X = 1$) iff Alice's original ball has the same color asthe second fusion product ($A = Y$). The second fusion product is green ($Y = 1$) iff Bob's original ball has the same color as the first fusion product ($B = X$). This reproduces equations 1.

As stipulated, this is a fine example. But one should be clear that these are no ordinary bowling balls, even leaving aside the fictional color fusion dynamics. Realistic bowling balls would involve a staggering number of degrees of freedom, realistically only describable through statistical means. Collectively, these degrees of freedom allow a much wider range of nomic possibilities than we could ever recognize here, including everything from spontaneous implosion, to disintegration, to spontaneous nuclear explosion. We could ignore these complexities and continue on with the bowling ball example. But I think it's better to avoid a false sense of security and instead work with clean toy examples. So I'll continue to focus my discussion around FORKS, leaving the bowling ball case as an option for those who reject a Lewisian miracle semantics.

I'll now turn to defending premises 3 – 5, taking them up in reverse order.

## 3 Defending *Non-Causation*

$A$ and $B$ occur on initially separate strands of spacetime. Up until they occur, and for a long time afterwards,[9] there is no spatiotemporal connection whatsoever between the two lines. As far as we know up to then, they could be forever spatiotemporally disconnected In the

[9]Unless otherwise indicated, any mention of "time" or duration refers to proper time.

far future, to be sure, the two lines converge. But what happens (say) a billion years out shouldn't matter for whether $A = 1$ causes $B = 1$.

We can enshrine these thoughts into a principle: what causes what within a history up to a time is intrinsic to that history. More precisely:

> **Weak Intrinsicality:** Let $w$ and $w'$ be nomologically compossible worlds with identical histories up to (and including / excluding) time $t$.[10] Then, for any propositions $\alpha, \beta$ purely about the history of $w$ up to (and including / excluding) time $t$, if it's true at $w$ that $\alpha$ causes $\beta$, then it's true at $w'$ that $\alpha$ causes $\beta$.

Others before me have defended similar principles. Hall (2004) proposes the following stronger principle:[11]

> **Intrinsicality (Hall):** Let world $w$ contain $S$, "a structure of [event propositions] consisting of [event proposition] $E$, together with all of its causes back to some earlier time $t$. ... Let $C$ be some [event proposition] in $S$ [disjoint] from $E$". Then, if $w'$ is nomologically compossible with $w$ and contains $S$, then $C$ causes $E$ at $w'$. (Cf. Hall 2004.)

Hall's intrinsicality principle is stronger because it merely requires that $w$ and $w'$ share a small subset of $E$'s history, namely $E$'s causes up to some prior time $t$.

Now, as Hall (2007) elsewhere points out, this principle produces awkward results in certain canonical scenarios, notably *Switching* and *Threat Cancelation*. Take Switching: If I activate a switch that controls which cable carries the current to the lamp, I thereby cause the current's flowing through (say) the left-hand rather than right-hand cable, but intuitively I haven't caused the *lamp's lighting up*. But now take a duplicate world, except that the right-hand cable is grounded instead of connected to the lamp. Here my flipping the switch *does* cause the lamp's lighting up. But the structure of the lamp, the left cable, and the switch exist in both worlds, and those arguably exhaust the relevant causes of the lamp's lighting up. So Hall's Intrinsicality principle would predict, wrongly, that my flipping the switch causes the lamp's lighting up even in the original world.

Threat Cancelation: I turn on a powerful electromagnet, and it deflects a piece of shrapnel that was hurling toward a window. Intuitively, the electromagnet's turning on is a cause of the window's staying intact. The causal past of the window's remaining intact consists, besides of the event itself, of the electromagnetic field's increase, and the electromagnet's

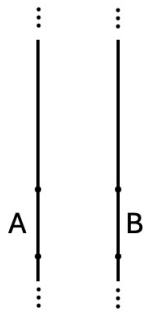
[10] "Identical", as in *numerical identity*—i.e., $w$ and $w'$ overlap up to $t$. One can also formulate a version of this principle in terms of qualitative duplication, which will be friendly to those who think that worlds don't overlap. That principle just introduces some needless complications.

[11] Where I've adjusted his formulation to the present framework, where the causal relata are propositions, worlds can overlap, and "disjointness" replaces Lewis's placeholder "distinctness".

turning on. But these three events are also present in a world where no piece of shrapnel is hurling toward the window to begin with. And in that world the electromagnet's turning on is causally unrelated to the window's remaining intact. So Hall's Intrinsicality principle would predict, wrongly, that the electromagnet's turning on isn't a cause of the window's remaining intact in the original world.

These cases threaten Hall's Intrinsicality principle, but they don't threaten **Weak Intrinsicality**. The temporal history of the lamp's lighting up entails how both cables are connected. Likewise, the temporal history of the window's remaining intact contains the piece of shrapnel.

It's easy to see how **Weak Intrinsicality** supports our initial reasoning. Consider TWO LINES, a world which shares FORK's history up to (though not including) the starting points of $X$ and $Z$. Thereafter the lines continue indefinitely without ever intersecting.




$A = 1$ obviously doesn't cause $B = 1$ in TWO LINES. But then it follows by **Weak Intrinsicality** that $A = 1$ doesn't cause $B = 1$ in FORKs.

# 4 Defending *Dependence*

As we'll see in section 7, **Dependence** is entailed by some leading structural equation accounts of causation, notably Halpern and Pearl (2005) and Halpern (2016). So it already has the blessing of some of the accounts we'll later critique.

Independently, **Dependence** follows from the conjunction of two claims:

(a) If it had been the case that $A = 0$, then the spacetime structure everywhere not upstream of $A$ would have been the same.

(b) If it had been the case that $A = 0$, then the dynamics everywhere not upstream of $A$ would have been the same.

By the identity rule (i.e., $\vdash \alpha \mathrel{\Box\!\!\rightarrow} \alpha$) and agglomeration,[12] (a) and (b) entail

[12]I.e. the rule $\vdash [(\alpha \mathrel{\Box\!\!\rightarrow} \beta) \wedge (\alpha \mathrel{\Box\!\!\rightarrow} \gamma)] \rightarrow [\alpha \mathrel{\Box\!\!\rightarrow} (\beta \wedge \gamma)]$, part of any standard logic of counterfactuals, including Lewis (1973b) and Stalnaker (1968).

(c) If it had been the case that $A = 0$, then $A = 0$ and, everywhere not upstream of $A$, the spacetime structure and the dynamics would be the same.

But (c)'s consequent logically entails that $B = 0$. Hence we have:

(d) If it had been the case that $A = 0$, then it would have been the case that $B = 0$.

Why believe premises (a) and (b)? Start with premise (b). It says that the dynamical laws are counterfactually robust, everywhere not upstream of $A$. So in particular it's entailed by any account of counterfactuals that entails laws' counterfactual robustness (*simpliciter*), e.g. Maudlin (2007) and Dorr (2016). It remains to convince those who deny laws' robustness.

Lewis's (1973; 1979) "miracles"-based account is the prime representative of this latter group. According to Lewis, if the actual laws are deterministic, then if things had been different, the actual laws would be violated—a "miracle", an event violating the actual laws, would have occurred. On Lewis's semantics (1979), provided there's a strict linear time order, miracles are thought to be confined to *before* the time of the antecedent. Now, Elga (2000) has shown (in my view quite conclusively) that Lewis's (1979) particular "hierarchy of importance" fails to produce the desired asymmetry of miracles. So in the following I won't focus on Lewis's hierarchy in particular, but instead grant that there's *some*—as yet unspecified—reductive semantics which produces the desired asymmetry of miracles.

In FORKS, the parts prior to $X$ and $Y$ have enough of a time order to make sense of an asymmetric "before": the intervals *before* $A$ are just those for which there's a future-directed causal curve running from them to $A$—i.e. just those intervals *upstream* of $A$. No such interval has additionally a past-directed causal curve running from it to $A$. Hence "before" is asymmetric, and so the asymmetry of miracles predicts that all miracles are confined to upstream of $A$. Hence the miracles-based view poses no threat to (b).

But perhaps we've construed the miracles asymmetry too strongly. Instead of confining miracles to *before* the antecedent, isn't it enough to merely confine them to *not after* the antecedent? After all, violating this stricture is what got Lewis (1973) originally into trouble: Fine (1975) famously notes that cheap *future* miracles are disastrous for Lewis. If future miracles were cheap, we'd have to conclude that, even if Nixon had pressed the nuclear button there'd be no nuclear war, since the effects of his pressing would have been nullified by a small future miracle. Fine's objection can already be avoided by deeming *future* miracles prohibitively costly.

But if we only avoid future miracles, then in addition to the pre-$A$ miracle in FORKS the closest $A = 0$-world could contain a *post-B* miracle, changing the field back to 0 after $B = 1$. The resulting world would witness $A = 0 \wedge B = 1$. But then premise (b) would fail and indeed we'd have $\neg(A = 0 \mathrel{\square\!\!\rightarrow} B = 0)$. The **Sufficiency** proponent could thus avoid the conclusion that $A = 1$ causes $B = 1$.

But it's hard to see a principled reason why the closest $A = 0$-world should include a miracle after $B$, short of simply wanting to ensure that $A = 1$ doesn't cause $B = 1$ according to **Sufficiency**. A miracle's job is to make the antecedent true. In FORKS that's possible without also breaking the law on the $B$-branch. Granted, miracles ought to bring about the antecedent with "minimal impact" to similarity, and perhaps there's something to be said about the world with a post-$B$ miracle being, in some pretheoretic sense, similar to the actual world (for example, it perfectly matches actuality on $B$'s branch prior to the miracle). But it's long been known that the relevant notion of similarity is a technical one, not tracking our pretheoretic similarity judgments.[13] So unless there's a principled theory of similarity vindicating those judgments, we shouldn't accept this reply

But even if there was such a theory, a moment's reflection reveals that we'd only push the problem back. For consider any interval $B'$ *between* the post-$B$ miracle and $Y$. We still want to say that $A = 1$ doesn't cause $B' = 1$, yet (by stipulation) there's no miracle that switches the field to 0 post-$B'$. This shows that ***Sufficiency*** *can't be saved by miracles confined to B's branch of* FORKS—at least one has to occur *downstream* from $A$. But *such* a miracle is supposed to be costly. At the very least, it's hard to see how any principled reason for placing a miracle downstream from $A$ in FORKS wouldn't also carry over to placing a miracle in the future of Nixon's pressing the button.

On to Premise (a). There are various ways one could manipulate the spacetime structure to make $A = 0$ compatible with $B = 1$. A minimally invasive approach might sever causal lines by deleting individual spacetime points. Specifically, to make $B = 1$ compatible with $A = 0$, one might either (i) sever the line connecting $B$ to $Y$ at some point prior to $Y$, or (ii) sever a line somewhere downstream from $A$. If the resulting world witnesses $A = 0 \wedge B = 1$ and is among the closest $A = 0$-worlds, then we'd have $\neg(A = 0 \mathrel{\square\!\!\rightarrow} B = 0)$. Again, the **Sufficiency** proponent could avoid the conclusion that $A = 1$ causes $B = 1$.

But those strategies have the same problems as the previous ones. Strategy (i) simply pushes the problem back: we can restate the problem for any interval $B'$ *between* the cut and $Y$. And strategy (ii) assumes that topology changes downstream from the antecedent come cheap. But if this was so, then we'd be back at the Nixon problem: his button press wouldn't have led to nuclear war because the signal in Nixon's cable would have (say) been swallowed by a small instantaneous singularity.

One might think that a third strategy to resist Premise (a) would be to sever $B$'s branch altogether from the rest of spacetime. This would convert the fork point at the start of $Y$ into a boring point. But this configuration is outright incompatible with $A = 0$: no solution to

[13] Lewis (1979, p. 466) himself notes that 'the similarity relation ... disagrees with ... explicit judgments of what is "very different"'. His argument is based on the observation, which he attributes independently to Pavel Tichý and Richard J. Hall, that "sometimes a pair of counterfactuals of the following form seem true: 'If A, the world would be very different; but if A and B, the world would not be very different.'" (ibid.).

the equations $X := Y \leftrightarrow A$ and $Y := X$ exists in which $A = 0$. So no $A = 0$-world with this configuration can be closest to actuality.

So I conclude that both premises stand. Now, in the introduction I emphasized the importance of non-backtracking readings for **Sufficiency**. Does the present argument somehow rely on interpreting one or more of the premises and conclusions according to a backtracking reading? No: premises (a) and (b) seem acceptable on a non-backtracking reading. We assent to them without reasoning backwards from the antecedent, unlike (say) with conditional (2), where our assent is predicated on an explicit inference about what must have preceded my not returning the book. But if premises (a) and (b) are acceptable on non-backtracking readings, then so must be the conclusions (c) and (d). According to Lewis (1979), whether a conditional has its backtracking or non-backtracking reading is fixed by context. But an argument's validity implies that its conclusion is acceptable in the same context as its premises. (For example, conjunction elimination entitles you to infer "Anupum is tall" from "Anupum is tall and Britt is tall" *if* and only if the conclusion is evaluated relative to the same standard of tallness as the premise.) Hence (c) and (d) are acceptable on a non-backtracking reading.

Considering (c) and (d) in isolation further supports this conclusion: there's no evidence that either of them is acceptable only on a backtracking reading, simply because no part of the $B$ branch is prior to $A$. (Now, no part of $B$ is *future* of $A$ either. But surely those counterfactuals can be non-backtracking—otherwise counterfactualist theories of causation would struggle to accommodate action at a distance.) To remove any ambiguity about this, we can also modify FORKS: first add additional temporal structure to FORKS, in the form of a privileged mapping from spacetime points into the real numbers that respects the manifold orientation. This mapping indicates how much time passes between any two spacetime points in FORKS. Given any such mapping, in at least one of the two initial branches there'll be an interval that's (wholly) later than some interval on the other branch. Label the earlier interval $A$ and the later interval $B$. Second, delete everything from the $B$-branch prior to $B$, and everything from the $A$-branch prior to $A$. What remains is a world in which the $A$-branch starts to exist, and some time later the $B$-branch starts to exist. But this change doesn't make it more plausible that $A = 1$ causes $B = 1$: the Weak Intrinsicality argument from the last section carries over without issue. But now it's entirely unambiguous that the counterfactuals (a) – (d), which all remain true, are non-backtracking. For there is simply nowhere to backtrack to—there's no

universe prior to $A$.[14]

# 5 Defending *Possibility*

**Possibility** asserts that FORKS is metaphysically possible. But you might resist this idea because you think that closed causal curves are metaphysically impossible. I'll first give positive reasons for their possibility, and afterwards address some possible concerns.

I take well-understood mathematical models of spacetimes to be powerful defeasible guides to metaphysical possibility. One popular way to cash this out is in terms of conceivability: to first approximation, spacetimes represented by well-understood mathematical models are conceivable, and conceivable things are possible. The inference from conceivability to possibility is an intuitive one, and suitably qualified forms a pillar of philosophical methodology, in the form of thought experiments.

As Chalmers (2002) points out, the situations in thought experiments are *positively* conceivable. Very roughly, negative conceivability involves merely an imagining of the absence of a contradiction in a proposition. Positive conceivability involves in addition an imagining of a "positive picture" of the proposed situation (Chalmers, 2002). A positive-conceivability-to-possibility link fends off well-known initial counterexamples to its naive counterpart: unprovable mathematical truths (such as Goldbach's conjecture perhaps) are metaphysically

[14]Both Lewis (1979) and Maudlin (2007) draw attention to the fact that for some counterfactual conditionals—to their mind, backtracking ones—replacing "would" by "would have to" preserves (and perhaps improves) felicity; as, for example, in the following case: "if I hadn't returned the book today, it would *have to have been* because Susy and I agreed on a later date to begin with." And performing the replacement on (d) does seem to preserve felicity:

(d*) If it had been the case that $A = 0$, then it would have to have been the case that $B = 0$.

Now, granted that (d) *is* acceptable on a non-backtracking reading, the test must be imperfect. Still, one might use the Lewis-Maudlin criterion to draw a different dividing line among counterfactuals, and then hope that this line separates *causal* and *non-causal* counterfactuals. But the following scenario suggests that the test doesn't separate causal from non-causal counterfactuals. Suppose you're skeptical that your child was home yesterday (and indeed she wasn't home). Your child claims to not have heard any loud noises yesterday, yet you know that your neighbor set off loud fireworks at night. You press your child repeatedly, yet he still denies hearing any loud bang. It then seems natural to insist: "If you had been home yesterday, then you *would have to have* heard some loud bangs." This counterfactual is causal: not being home caused his not hearing the bang. So the Lewis-Maudlin test doesn't separate causal from non-causal counterfactuals.

So if it distinguishes neither backtracking from non-backtracking, nor causal from non-causal, counterfactuals, what *does* the test measure? One hypothesis is that "have to" is an epistemic modal, used to indicate that the speaker's evidential support for a statement is *indirect* (Mandelkern, 2019), e.g. that the statement is inferred from some implicit premise. For example, "It has to be raining" sounds strange if the speaker directly observes the rain; but it sounds *good* if she infers it indirectly, e.g. from the fact that water is leaking through her roof. Hence also why the earlier conditional about your child's whereabouts sounded felicitous: it emphasized that you inferred that your child would have heard a loud bang from the fact that your neighbor set off fireworks yesterday. Perhaps the modal plays the same role in "If the moon were to explode right now, our scientists would have to have known about it before": it emphasizes that the consequent is inferred from some implicit premise—e.g. that our scientists have great foresight.

necessary, and their falsity arguably negatively, but not positively, conceivable. Similarly, after learning the relevant empirical facts, the falsity of empirical necessities (e.g. that water is $H_2O$) are perhaps negatively conceivable, but arguably not positively conceivable.

Worlds that are entirely represented by fully interpreted mathematical models—like FORKS—are paradigms of positive conceivability: they present a precise and detailed positive picture. Now, "fully interpreted" is doing work here: where it's controversial what a given aspect of a mathematical model represents, it'll generally be controversial what's metaphysically possible. Consider the debate about haecceitism in the metaphysics of spacetime: do swaps of mathematical points correspond to possible swaps of spacetime points?[15] Fortunately, the representational aspects of FORKS we are sensitive to are all fully interpreted. For example, it's uncontroversial that the directed mathematical line from the starting point of the mathematical interval representing $X$ back to itself represents a closed time-like path in spacetime. These sorts of facts are all we need.

In addition to considering conceivability in the abstract, we also shouldn't forget that the Einstein equations themselves have exact solutions with closed causal curves. In Kurt Gödel's (1949) example (which, remarkably, is homeomorphic to $\mathbb{R}^4$), suitably accelerated material bodies can travel along closed time-like curves. Other examples of spacetimes with closed causal curves include ones with rotating black holes (Kerr solutions, cf. Carter (1968)) and Van Stockum's rotating dust cylinders (Stockum, 1938).[16] Ordinarily, one would accuse those of philosophical overreach who banish, on *a priori* grounds, a substantial part of the scientific literature to the metaphysically impossible.

Of course, the charge of philosophical overreach can be countered by strong positive arguments for the philosophical conclusion. Are there any for the impossibility of causal loops? One of the more influential worries stems from cognates of the grandfather paradox. Let *autoinfanticide* be the act of a future self's (permanently) killing her own infant self. You can't possibly commit autoinfanticide. But if causal loops are possible, then (it seems) you *can* possibly commit autoinfanticide: travel via a causal loop back in time, and position yourself in front of your own crib, gun in hand. Here and now, you have what it takes: you are a good shot, you can pull the trigger, etc. It would thus seem that, if causal loops are possible, then it's both the case and not the case that you can possibly commit autoinfanticide—contradiction. So causal loops aren't possible.

I find the standard reply to this, due to Lewis (1976), convincing: what you "can" do is highly context-sensitive. To quote Lewis's example: compared to a (non-human) ape, I *can* speak Finnish: I have sufficiently developed articulators. But compared to a Finnish speaker, I *can't* speak Finnish: I don't know any Finnish vocabulary or grammar. Following Lewis, we

[15] Or as the question is often put: do diffeomorphically equivalent Lorentzian manifolds represent "genuinely distinct" possibilities? See e.g. Norton, Pooley, and Read (2023) for an overview of the debate.

[16] See also Kajari et al. (2004).

may say that a speaker $S$ can $\phi$ in context $C$ iff $S$'s $\phi$-ing is metaphysically compossible with $C$.

There are plenty of contexts in which I *can* kill the baby in the crate: e.g. most contexts which leave out that the baby is me. They witness the fact that "I have what it takes". They don't imply, however, that I can commit *autoinfanticide*. So the apparent contradiction is resolved.

Another less common family of objections concerns the "bootstrapping" aspect of causal loops. Here I'll just mention one member of the family (for the rest, and rebuttals against them, see e.g. Effingham (2020, Ch. 5.2.2)). The objection is that causal loops are inexplicable (Al-Khalili, 1999). But are inexplicable things impossible? No, as Lewis (1976, p. 148) already notes: the universe's entire past is plausibly *actually* inexplicable; or if it isn't, it at least *possibly* is. And so is God. And so are outcomes of genuinely stochastic processes. So inexplicability doesn't entail impossibility.

In summary, I conclude that causal loops are metaphysically possible. But, as far as worlds with causal loops are concerned, there is nothing special about FORKS. So FORKS is metaphysically possible.

# 6 Troubles for Accounts of Causation

## 6.1 Against Lewis (1973a) and Hall (2007)

So, there can be determinate, non-causal counterfactual dependence, even between positive, proportional, and macroscopic events. What does that mean for counterfactualist reductions of causation? Most obviously, any account which entails **Sufficiency** will have to be rejected. At least two such accounts come to mind.

The most famous is Lewis (1973a). Translated into our ideology of event propositions, it says that $A$ causes $B$ iff there is a chain of true event propositions $X_1, ..., X_n$ with $X_1 = A$ and $X_n = B$ such that, for all $i = 1, ..., n-1$, $X_i$ and $X_{i+1}$ are disjoint and $\neg X_i \boxbslash\!\!\rightarrow \neg X_{i+1}$. In particular, then, $\neg A \boxbslash\!\!\rightarrow \neg B$ with disjoint $A$ and $B$ is sufficient for $A$'s causing $B$. Hence Lewis's account should be rejected. Our argument joins the ranks of many previous objections raised against Lewis—notably his account's failure to handle cases of late preemption and symmetric overdetermination. But in contrast to those objections, our argument also applies to successor theories.

One of those successors is Hall (2007). After a forceful critique of extant structural equations accounts of causation (more on those below), Hall puts forth his own account. According to it, $A = n$ causes $B = m$ iff $A = n$ and $B = m$ are true and there is a "*reduction*" of the actual situation in which $B = m$ counterfactually depends on $A = n$. It needn't concern

us what exactly a reduction is. (Just to give a flavor: roughly, it's a situation in which zero or more parts that are actually in a "non-default" state adopt their default state instead, and the rest is unchanged.) What matters here is that every situation counts as a reduction of *itself* (cf. also Hall (2007, p. 127)). Thus we have, again, that $\neg A \mathrel{\Box\!\!\rightarrow} \neg B$ is sufficient for $A$'s causing $B$. So Hall's (2007) account must be rejected too.[17,18]

## 6.2 Against Halpern and Pearl (2005), and Cognates

Halpern and Pearl (2005) is an early and influential structural-equation-based analysis of causation ("seminal", according to Beckers and Vennekens (2017)). It is recently endorsed also in Halpern (2016). Notably, the account is explicitly developed with cyclic (or "non-recursive") causal structures in mind. Unfortunately, it yields the wrong results for mild variants of FORKS.

Halpern and Pearl provide a model-relative definition of "cause". Roughly, a causal model is a coarse-grained representation of the world's causal structure. It uses variables to represent sets of events and structural equations to represent dependencies between these events. Formally, it is a triple $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathcal{E})$ consisting of a set $\mathbf{U}$ of *exogenous* variables, a set $\mathbf{V}$ of *endogenous* variables, and a set $\mathcal{E}$ of structural equations in members of $\mathbf{U} \cup \mathbf{V}$. By definition, *exogenous* variable don't appear anywhere on the left-hand sides of the structural equations, and every *endogenous* variable appears on the left-hand side of exactly one of the structural equations. A *context* for $\mathcal{M}$ is an assignment of values to the variables in $\mathbf{U}$. (Henceforth, I'll use **bold-face** to indicate sets of variables and *light-face* to indicate individual variables.)

To first approximation, Halpern and Pearl want to capture the idea that $X = x$ is an actual cause of $Y = y$ in a given model iff $X = x$ and $Y = y$ are both true and the model contains some relevant actual or non-actual circumstances ("contingencies") in which $Y = y$ counterfactually depends on $X = x$. The relevant circumstances are given by states of the surrounding variables; more specifically, according to Halpern and Pearl, they are exactly those states which don't already by themselves force a change in $Y$'s value.

More formally, and to second approximation, $X = x$ is an actual cause of $Y = y$, relative

[17] To Hall's defense, he is aware of the limitations of his account, explicitly bracketing the case of causal loops (p. 114, esp. fn. 6). But this doesn't change the fact that his account isn't a satisfactory analysis of causation.

[18] Earlier I discussed Glynn's (2013) account (cf. fn. 4). It's easy to see that it, too, succumbs to the present counterexample. Assume, for the account's sake, that FORKS is imbued with additional temporal structure, identifying times across the two branches. If we insist that in evaluating $\neg A \mathrel{\Box\!\!\rightarrow} \neg B$ we keep everything prior to $A$ as it actually is, a miracle must occur on the $B$-branch sometime after $t$. Suppose that the miracle occurs at $t^+$.

We then get the result that, while the $A$-branch's value at $t$ isn't a cause of the $B$-branch's value before $t$, it *is* a cause of the $B$-branch's value *after* $t^+$. But that seems wrong—instead we should say that the $B$-branch's value at $t^+$ is fully caused by *its* past values, with the $A$-branch having no causal influence.

to a causal model $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathcal{E})$ in context $\mathbf{U} = \mathbf{u}$, iff $X = x$ and $Y = y$ are true in $\mathcal{M}$ and there are a set of endogenous variables $\mathbf{W} \subseteq \mathbf{V} \backslash \{X, Y\}$ and values $\mathbf{w}'$ such that, in the model resulting from setting $\mathbf{W} = \mathbf{w}'$ and holding all exogenous variables $\mathbf{U}$ fixed at $\mathbf{u}$, $Y = y$ is still true but now depends on $X = x$. That is, for some $x' \neq x$, the structural equations resulting from manually fixing $\mathbf{U} = \mathbf{u}$, $\mathbf{W} = \mathbf{w}'$, and $X = x$, entail a different value for $Y$ than the structural equations resulting from manually fixing $\mathbf{U} = \mathbf{u}$, $\mathbf{W} = \mathbf{w}'$, and $X = x'$.[19,20] Note, in particular, that in either case Halpern and Pearl hold fixed all exogenous variables at the given context.[21] This turns out to save the account's predictions in the case of FORKS.

[19] This approximate description is enough to make our case. But for completeness, here is the precise formulation of Halpern and Pearl's account. Let a *primitive event* be any sentence of the form $X = x$ for $X \in \mathbf{U}$. Where $\varphi$ is a Boolean combination of primitive events in $\mathcal{M}$, say that $\varphi$ is true in $(\mathcal{M}, \mathbf{u}, \mathbf{v})$ iff $\mathbf{V} = \mathbf{v}$ entails $\varphi$. Moreover, for any $\mathbf{Y} \subseteq \mathbf{V}$, say that the sentence $[\mathbf{Y} \leftarrow \mathbf{y}]\varphi$ is true in $(\mathcal{M}, \mathbf{u}, \mathbf{v})$ iff $\varphi$ is true in all $(\mathcal{M}_{\mathbf{Y} \leftarrow \mathbf{y}}, \mathbf{u}, \mathbf{v}')$, where $\mathcal{M}_{\mathbf{Y} \leftarrow \mathbf{y}}$ is the result of replacing in $\mathcal{M}$ the right-hand side of any structural equation for a variable in $\mathbf{Y}$ with the respective value in $\mathbf{y}$, and $(\mathbf{u}, \mathbf{v}')$ is a solution to the resulting structural equations. $\langle \mathbf{Y} \leftarrow \mathbf{y} \rangle \varphi$ abbreviates $\neg ([\mathbf{Y} \leftarrow \mathbf{y}] \neg \varphi)$. Then:

> "$\mathbf{X} = \mathbf{x}$ is an *actual cause* of $\varphi$ in $(M, \mathbf{u}, \mathbf{v})$ if[f] the following three conditions hold:
>
> AC1. $(M, \mathbf{u}, \mathbf{v}) \vDash (\mathbf{X} = \mathbf{x}) \wedge \varphi$.
>
> AC2. There exists a partition $(\mathbf{Z}, \mathbf{W})$ of $\mathbf{V}$ with $\mathbf{X} \subseteq \mathbf{Z}$ and some setting $(\mathbf{x}', \mathbf{w}')$ of the variables in $(\mathbf{X}, \mathbf{W})$ such that if $(M, \mathbf{u}, \mathbf{v}) \vDash (\mathbf{Z} = \mathbf{z}^*)$, then the following conditions hold:
>
> (a) $(M, \mathbf{u}, \mathbf{v}) \vDash \langle \mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}' \rangle \neg \varphi$.
> (b) $(M, \mathbf{u}, \mathbf{v}) \vDash [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{W} \leftarrow \mathbf{w}', \mathbf{Z}' \leftarrow \mathbf{z}^*] \varphi$ for all subsets $\mathbf{Z}'$ of $\mathbf{Z}$.
>
> AC3. $\mathbf{X}$ is minimal; no [proper] subset of $\mathbf{X}$ satisfies conditions AC1 and AC2." (Halpern and Pearl, 2005, p. 884, notation adjusted)

(Oddly, this definition generalizes Halpern and Pearl's "original", *pre*-2005 definition of causation (cf. Halpern, 2000). To avoid counterexamples discovered in the interim, Halpern and Pearl (2005) replace, in the *acyclic* definition, AC2.(b) with the following condition, which quantifies additionally over all subsets of $\mathbf{W}$ (Halpern (2016, Ch. 2.8) calls this the "updated" definition):

> AC2(b)*. $(M, \mathbf{u}, \mathbf{v}) \vDash [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{W}' \leftarrow \mathbf{w}', \mathbf{Z}' \leftarrow \mathbf{z}^*] \varphi$ for all subsets $\mathbf{W}'$ of $\mathbf{W}$ and $\mathbf{Z}'$ of $\mathbf{Z}$.

Halpern and Pearl's using the old condition for the cyclic case is presumably a simple oversight. After all, the cyclic definition is supposed to *generalize* the acyclic definition. Indeed, Halpern (2016, p. 57) gives the correct cyclic formulation, with AC2.(b)* replacing AC2.(b). In any case, this makes no difference to us, because in our example $\mathbf{W} = \emptyset$ anyway.)

[20] Halpern (2016) discusses, besides the "original" and the "updated" definition (see fn. 19), a third one, which he calls the "modified" definition. This definition replaces the whole of AC2 with a simpler condition, which only considers the actual values of $\mathbf{W}$ (see Halpern (2016, p. 25, 57)):

> AC2(m). There is a set $\mathbf{W}$ of variables in $\mathcal{V}$ and a setting $\mathbf{x}'$ of the variables in $\mathbf{X}$ such that if $(M, \mathbf{u}) \vDash \mathbf{W} \leftarrow \mathbf{w}^*$, then
>
> $$(M, \mathbf{u}) \vDash \langle X \leftarrow x', \mathbf{W} \leftarrow \mathbf{w}^* \rangle \neg \varphi.$$

Again, this change doesn't matter for us, since $\mathbf{W} = \emptyset$ in our example.

[21] Curiously, by holding *all* exogenous variables fixed, Halpern and Pearl can only provide a theory of when *endogenous* variables are causes (cf. also fn. 19). This is against their own assertion (see their p. 847)! The obvious fix is to permit $X \in \mathbf{U}$, and to hold only the *non-X* exogenous variables fixed.

Before we get there, a methodological remark. As others have pointed out (e.g. Hall (2007)), we can't stop at a purely model-relative account of causation. My letting go of the pen didn't just cause it to fall relative to one or another model—it caused it to fall *full stop*. Any serious account of causation should reproduce this judgment. The causal modeling literature therefore introduces the notion of a model's *adequacy*: $X = x$ causes $Y = y$ *simpliciter* iff $X = x$ causes $Y = y$ relative to an *adequate* causal model.

The nomic structural equations in eqs. 1 naturally constitute an adequate model of FORKS. Now, Halpern and Pearl say little about adequacy, but what they do say lines up: according to them, adequate models encode "generic causal knowledge such as *what we obtain from the equations of physics*" (p. 849, my emphasis)—our nomic structural equations encode exactly those equations. In my discussion of Halpern and Pearl, I'll therefore assume that the equations in eqs. 1 are an adequate causal model for FORKS.

To evaluate whether $A = 1$ causes $B = 1$ according to Halpern and Pearl, we add two exogenous variables, $A^*$ and $B^*$, to the model representing intervals preceding $A$ and $B$, respectively. (This is needed because Halpern and Pearl only define causation as a relation between endogenous variables—cf. 21.) We also add the following structural equations to eqs. 1:

$$A := A^*,$$
$$B := B^*.$$

The resulting model is clearly still adequate.

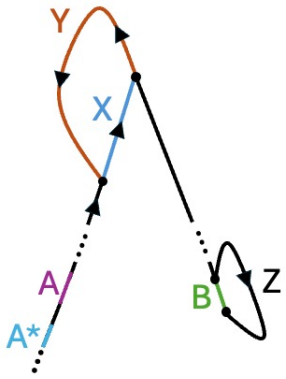
Holding fixed $B^* = 1$, $B = 1$ no longer counterfactually depends on $A = 1$.[22] So, it turns out, Halpern and Pearl's account correctly judges that $A = 1$ doesn't cause $B = 1$.[23] Moreover, as the reader may verify (with the help of fn. 19), Halpern and Pearl's account also correctly implies that $A$ and $Y$ each cause $X$ and that $X$ and $B$ each cause $Y$.

(Since $B = 1$ counterfactually depends on $A = 1$, yet Halpern and Pearl's account doesn't entail that $A = 1$ causes $B = 1$, we see that Beckers and Vennekens's (2017) quote in section 1 is mistaken: Halpern and Pearl's (2005) account doesn't entail **Sufficiency**—indeed, it's incompatible with it. We'll see that all of the cited accounts are, except for Lewis (1973a) and Hall (2007).)

[22] More precisely (to put it in terms of fn. 19): $\mathcal{M}_{A\leftarrow 0}$ doesn't have a solution with $B^* = 1$. Hence, *vacuously*, $(\mathcal{M}, A^* = B^* = 1, \mathbf{v}) \vDash [A \leftarrow 0](B = 1)$, and so condition AC2.(a) is violated for $\mathbf{W} = \emptyset$. Moreover, for any non-empty $\mathbf{W} \subseteq \{X, Y\}$, $\mathcal{M}_{\mathbf{W}\leftarrow\mathbf{w}}$ is loop-free, and it's easy to verify that in such a model $A$ and $B$ are counterfactually independent. So we have $(\mathcal{M}, A^* = B^* = 1, \mathbf{v}) \vDash [A \leftarrow 0, \mathbf{W} \leftarrow \mathbf{w}'](B = 1)$ for any $\mathbf{W} \subseteq \{X, Y\}$. Finally, whenever $B \in \mathbf{W}$, AC2.(a) is violated if $\mathbf{w}'$ assigns 1 to $B$ and AC2.(b) is violated otherwise. So either AC2.(a) or AC2.(b) is violated for any $\mathbf{W} \subseteq \mathbf{V}\backslash\{A\}$ and any values $\mathbf{w}'$ of $\mathbf{W}$.

[23] At least relative to this causal model. In principle it's possible that $A = 1$ causes $B = 1$ relative to some other adequate causal model (in which $A$ and $B$ are endogenous). But, for the sake of argument, let's grant Halpern and Pearl that this is not the case.

However, Halpern and Pearl's success crucially depends on $B$'s value being fixed by exogenous variables. This is an artifact of the specific case of FORKS. Consider a spacetime in which, additionally, *B itself* is on a loop. For example, consider MOD-FORKS, a world with added region $Z$, as follows:




where the equations in 1 are extended by the following three equations:

$$A := A^*,$$
$$B := Z,$$
$$Z := B.$$

Clearly, $A = 1$ still doesn't cause $B = 1$. But note that in this new model, neither $B$ *nor any of its ancestors* are exogenous. As a result, even when we hold all exogenous variables fixed (setting $\mathbf{W} = \emptyset$), $B = 1$ counterfactually depends on $A = 1$. Hence Halpern and Pearl's account implies that $A = 1$ causes $B = 1$ relative to this model.[24]

(Halpern and Pearl (2005) also allow models with *infinitely* many variables (requiring only minor modifications to their proposal which won't matter here—see their pp. 883-4). Hence we can achieve the same result even without switching worlds. For consider a model of FORKS where $B$ has infinitely many ancestors, $Z_1, Z_2, Z_3, \ldots$ representing disjoint intervals jointly reaching infinitely into $B$'s past. The structural equations added to 1 are as follows:

$$A := A^*,$$
$$B := Z_1,$$
$$Z_1 := Z_2,$$
$$Z_2 := Z_3,$$
$$\ldots$$

[24]More precisely (to put it in terms of fn. 19): the triple $(\mathcal{M}, A^* = 1, A = X = Y = B = Z = 1)$ entails both $[A \leftarrow 0]B = 0$ (satisfying AC2a) and $[A \leftarrow 1, \mathbf{Z} \leftarrow \mathbf{1}]B = 1$ for any $\mathbf{Z} \subseteq \{X, Y, Z, B\}$ (satisfying AC2b). Conditions AC1 and AC3 are immediate.

Again, $A = 1$ is still not a cause of $B = 1$. But neither $B$ *nor any of its ancestors* are exogenous. So, again, Halpern and Pearl's (2005) account implies wrongly that $A = 1$ causes $B = 1$.)

Some accounts copy enough from Halpern and Pearl (2005) to be vulnerable to the same objection. Beckers and Vennekens (2017, p. 5) hold that $C = c$ causes $E = e$ relative to a causal model and context $\mathbf{u}$, if the model resulting from replacing $C$'s structural equation with $C := c'$ (where $c' \neq c$) entails $E = e$ in context $\mathbf{u}$.[25] This straightforwardly implies that $A = 1$ causes $B = 1$ in MOD-FORKS.

*

The accounts thus far all fail for the same reason: they don't make the existence of a directed path from cause to effect necessary for causation. According to Lewis (1973a) and Hall (2007), counterfactual dependence suffices for causation, irrespective of whether there's such a path. For Halpern and Pearl, what matters for causation is that there are suitable global solutions to the structural equations. Their failure to make directed paths necessary for causation has largely gone unnoticed. For example, Hall (2007), Glynn (2013, p. 46-7), and Weslake (2015) all claim to reproduce Halpern and Pearl's (2005) account, but in fact describe accounts which make directed paths necessary for causation.[26] As MOD-FORKS shows, that's false.

(Interestingly, Halpern (2016) ostensibly *proves* that Halpern and Pearl's account implies the necessity of directed paths for causation.[27] Our discussion shows that the theorem is false—hence the proof is invalid. I present Halpern's theorem and diagnose its proof's invalidity in the following footnote.[28] Note that an analogous proof of a weaker version of

[25] Like Hall (2007), Beckers and Vennekens explicitly bracket the cyclic case (see their p. 3). But, again, we can't afford this if our goal is to provide a *definition* of causation.

[26] Weslake does clarify in a footnote: "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." FORKS is a concrete case where Weslake's reconstruction makes different predictions than Halpern and Pearl's account.

[27] This follows from the following "theorem" of Halpern's (for the meaning of "original" and "modified" definition, see fns. 19 and 20):

> "**Proposition 2.9.2** If $\mathbf{X} = \mathbf{x}$ is a cause of $\varphi$ in $(M, \mathbf{u})$ according to the original or modified [Halpern-Pearl] definition, then there is [a $\mathbf{W} \subseteq \mathcal{V} \backslash \mathbf{X}$] such that every variable $Z \in \mathcal{V} - \mathbf{W}$ lies on a causal path in $(M, \mathbf{u})$ from some variable in $\mathbf{X}$ to some variable in $\varphi$."

(For simplicity I've substituted $\mathbf{W} \subseteq \mathcal{V} \backslash \mathbf{X}$ for "a witness $(\mathbf{W}, \mathbf{w}, \mathbf{x}')$"; the $\mathbf{w}$ and $\mathbf{x}'$ don't play a role in the theorem. To see that witnesshood implies $\mathbf{W} \subseteq \mathcal{V} \backslash \mathbf{X}$, see Halpern's (2016, p. 25) definition of "witness".) To see how this entails the claim: since $\mathbf{X} \subseteq \mathcal{V} - \mathbf{W}$, then if $\mathbf{X}$ is non-empty, there's at least one $Z \in \mathcal{V} - \mathbf{W}$. Hence Proposition 2.9.2 entails that if there is a variable $X$ such that $X = x$ is a cause of $\varphi$ in (M, $\mathbf{u}$), then there exists a $Z$ which lies on a causal path from $X$ to some variable in $\varphi$. *A fortiori*, there is causal path from $X$ to some variable in $\varphi$.

[28] The first lemma:

> "**Lemma 2.10.1** If $Y$ and all the variables in $\mathbf{X}$ are endogenous, $Y \notin \mathbf{X}$ , and there is no causal path in $(M, \mathbf{u})$ from a variable in $\mathbf{X}$ to $Y$, then for all sets $\mathbf{W}$ of variables disjoint from $\mathbf{X}$ and $Y$, and all settings $\mathbf{x}$ and $\mathbf{x}'$ for $\mathbf{X}$, $\mathbf{y}$ for $\mathbf{Y}$, and $\mathbf{w}$ for $\mathbf{W}$, we have
>
> $$(M, \mathbf{u}) \vDash [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{W} \leftarrow \mathbf{w}](Y = y) \text{ iff } (M, \mathbf{u}) \vDash [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}](Y = y).$$

Halpern's theorem, restricted to acyclic models, would be valid.)

Our discussion suggests an obvious fix for our troubles: build in path-dependence from the start. That is, add that $X = x$ causes $Y = y$ *only if there's a directed path from* $X$ *to* $Y$. Indeed, this is what Hitchcock (2001), Menzies (2004), and Weslake (2015) do. Crucially, whether this strategy is successful still depends on having the right account of adequacy. If some "adequate" causal model of FORKS contains an edge from $A$ to $B$, then nothing is gained with a path-dependent analysis. So far, I've simply granted Halpern and Pearl that eqs. 1—which *don't* include an edge from $A$ to $B$—constitute an adequate causal model of MOD-FORKS. But a satisfactory reductive definition of causation should *derive* this result, not stipulate it. Let's see what others have said about adequacy.

# 7 Troubles for Accounts of Structural Equations

Minimally, an adequate causal model must contain only true (or at least approximately true) structural equations (cf. Hitchcock (2001, p. 292)). Hitchcock (2001), Menzies (2004), and Weslake (2015) broadly agree that the relevant truth conditions are given in terms of *counterfactuals*. Here is Hitchcock (2001, p. 280) (see also Hitchcock (2007, p. 500)):

> "[S]tructural equations encode counterfactuals. For example, $[Z = f_Z(X, Y, ..., W)]$ encodes a set of counterfactuals of the following form:
>
> > If it were the case that $X = x$, $Y = y$, ... , $W = w$, then it would be the case that $Z = f_Z(x, y, ..., w)$."

. . . " (Halpern, 2016, p. 66)

In other words, the lemma says that if there's no path from $\mathbf{X}$ to $Y$, then $Y$'s value doesn't depend on $X$'s value no matter if or how we set the other endogenous variables. But this lemma is invalid (though its weakening to the acyclic case is valid): in MOD-FORKS, there is no causal path from $A$ to $B$, yet $B$'s value depends on $A$'s value, e.g. $(M, A^* = 1) \models [A \leftarrow 0](B = 0)$ and $(M, A^* = 1) \not\models [A \leftarrow 1](B = 0)$. Halpern's proof proceeds by induction on the maximum distance of a variable to any exogenous variable. But in cyclic graphs a node generally doesn't have a finite maximum distance to any exogenous variable, hence the induction proof fails.

The second lemma:

> "**Lemma 2.10.2** If $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}](\mathbf{Y} = \mathbf{y})$, then $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}]\varphi$ if and only if $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{Y} \leftarrow \mathbf{y}]\varphi$." (Halpern, 2016, p. 68)

In other words, the lemma says that if setting $\mathbf{X} = \mathbf{x}$ entails $\mathbf{Y} = \mathbf{y}$ in a model, then setting $\mathbf{X} = \mathbf{x}$ entails $\varphi$ just in case setting both $\mathbf{X} = \mathbf{x}$ *and* $\mathbf{Y} = \mathbf{y}$ entails $\varphi$. To see that this is invalid, let $B'$ be an interval (disjoint from $Z$) on the way from $B$ to $Y$ in MOD-FORKS. Then $(M, A^* = 1) \models [A \leftarrow 1](B' = 1)$ and $(M, A^* = 1) \models [A \leftarrow 1](B = 1)$, but *not* $(M, A^* = 1) \models [A \leftarrow 1, B' \leftarrow 1](B = 1)$, since the model $M_{A \leftarrow 1, B' \leftarrow 1}$ has more than one solution, one with $B = Z = 1$ and another with $B = Z = 0$. Halpern's proof of this lemma presupposes that a model has a unique solution given an assignment of values to its exogenous variables, which (as he himself notes in Ch. 2.7) is generally false for cyclic models.

Similarly, Menzies (2004, p. 822):[29]

> "[The equation $SH = ST$] asserts that if Suzy threw a rock, her rock [would] hit the bottle; and if she didn't throw a rock, her rock [wouldn't have] hit the bottle."

Finally, Weslake (2015):

> "A causal model is a representational device for encoding counterfactual relationships between variables. Counterfactual relationships are represented by [structural] equations."

Of course, not all systems of true structural equations constitute adequate causal models. An adequate causal model must also be in some sense "non-redundant" or "minimal".[30] Of the three, only Hitchcock provides explicit non-redundancy conditions:

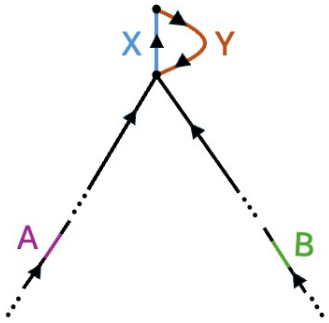
> "[For a model to be adequate, structural equations] must always be written in minimal form: [if, for every combination of values of $Y, ..., W$,] the value of $Z$ does not depend upon the value of $X$ ... [then] the structural equation for $Z$ [i.e., $Z = f_Z(X, Y, ..., W)$] must be rewritten $Z = f_Z(Y, ..., W)$. ... By the same token, [structural equations in an adequate causal model] must always include as arguments any variables ... upon which $Z$ counterfactually depends, given [some] values of the other variables." (p. 280-1)

Hitchcock's recipe involves evaluating counterfactuals whose antecedents specify arbitrary combinations of values for a model's variables. Some of those counterfactuals explicitly suspend the underlying causal structure. For example, suppose $X = x$ in fact causes $Y = y$ and we'd like to assess the adequacy of a structural equation with exactly $X$ and $Y$ on its right-hand side. We'll then have to evaluate the counterfactual "If $X = x$, but $Y \neq y$ anyway, then..."; the antecedent explicitly suspends the causal structure downstream from $X$. Hitchcock (2001, p. 275) calls these counterfactuals "*explicitly nonforetracking*" (ENF) counterfactuals.

Cyclic models reveal that Hitchcock's recipe is inadequate: it qualifies too many models as adequate. Consider the following relative of FORKS:

[29] Curiously, Menzies uses indicative conditionals here, even though he means them to be "counterfactuals", following "Halpern and Pearl". I've added the subjunctive form.

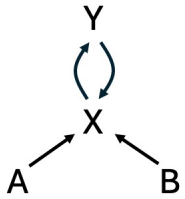
[30] Additionally there are constraints on variable choice. Minimally, an adequate model's variables should be "distinct", which throughout we've understood in terms of disjointness of event propositions. Halpern and Hitchcock (2010, Sec. 3) add two additional constraints: the first demands that a model's causal predictions be stable under introduction of new (distinct) variables, and the second demands that variable values be proportional to each other. The simple example below meets all three constraints; so in the following we'll keep them in the background.




Suppose that the dynamics at the three-way fork point dictate that, whenever $A$ and $B$ agree, then $X$ and $Y$ agree, and whenever $A$ and $B$ disagree, $X$ and $Y$ disagree. The rest is as usual. The here dynamics can be captured by the following nomic structural equations:

$$\begin{aligned} X &:= (A \leftrightarrow B) \cdot Y + (A \leftrightarrow \overline{B}) \cdot \overline{Y}, \\ Y &:= X. \end{aligned} \tag{2}$$

Suppose that actually $A = B = 1$ and $X = Y = 0$. Call this world THREE-WAY FORK. For illustration, the equations' graphical structure is this:




Like in FORKS, $A = 1$ plausibly doesn't cause $B = 1$. And like in FORKS, the dynamics again have exactly four solutions, now as follows:

| $A$ | $B$ | $X$ | $Y$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 1 |

That is, $A$ and $B$ must always agree, and $X$ and $Y$ must always agree; otherwise there are no constraints.

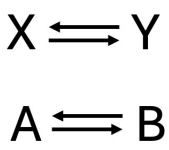
According to Hitchcock's criterion, in any adequate causal model for THREE-WAY FORK $B$'s structural equation must contain $A$. For the dynamics imply the truth of the following counterfactuals, for $z = 0, 1$:

$$A = 0 \wedge X = z \wedge Y = z \mathrel{\Box\!\!\rightarrow} B = 0,$$
$$A = 1 \wedge X = z \wedge Y = z \mathrel{\Box\!\!\rightarrow} B = 1. \tag{3}$$

So, there are values of the rest of the graph such that $B$ counterfactually depends on $A$, and so, by Hitchcock's recipe, $B$'s structural equation contains $A$, i.e.

$$f_B \equiv f_B(A, ...).$$

By similar reasoning, $A$'s structural equation must contain $B$, $X$'s structural equation must contain $Y$, and $Y$'s structural equation must contain $X$. Hence, according to Hitchcock, any adequate causal model contains an edge from $B$ to $A$. Indeed, it contains at least all of the following edges:[31]




Now, the definitions of causation in Hitchcock (2001), Menzies (2004), and Weslake (2015) entail the following "path-sensitive" variant of the sufficiency claim:

> **Sufficiency*:** Let $\mathcal{M}$ be an adequate causal model whose variables include $X$ and $Y$. Then $X = x$ causes $Y = y$ (*simpliciter*) if $\mathcal{M}$ contains a directed path $p$ from $X$ to $Y$ such that, when all variables not on the path are held fixed at their actual values, $Y = y$ counterfactually depends on $X = x$.

But $\mathcal{M}^*$ contains a directed path from $A$ to $B$ and moreover, as we see from eqs. 3, whenever $X$ and $Y$ are both held fixed at their actual values (either both 0 or both 1), $B = 1$ counterfactually depends on $A = 1$. So, given Hitchcock's criterion of adequacy, the definitions of causation in Hitchcock (2001), Menzies (2004), and Weslake (2015) all entail that $A = 1$ causes $B = 1$ in THREE-WAY FORK. But $A = 1$ doesn't cause $B = 1$ in THREE-WAY FORK. So, unless a different

[31] It's not entirely obvious whether there should be additional edges. For example, according to Hitchcock, $B$'s structural equation contains $X$ iff one of the following counterfactuals is true (for $x, y, z, x', z' \in \{0, 1\}$, $x \neq x'$, and $z \neq z'$):

$$A = z \wedge X = x \wedge Y = y \mathrel{\Box\!\!\rightarrow} B = z,$$
$$A = z \wedge X = x' \wedge Y = y \mathrel{\Box\!\!\rightarrow} B = z'.$$

Since one of $X = x \wedge Y = y$ and $X = x' \wedge Y = y$ is prohibited by the conjunction of spacetime structure and law, the laws and/or spacetime structure have to break in at least one counterfactual circumstance. How they break is, *prima facie*, anyone's guess. It won't matter for us.

account of adequacy is supplied, all three accounts should be rejected.[32]

# 8 Conclusion

I've argued that even weak versions of **Sufficiency** are false. If I'm right, this excludes some proposed counterfactual analyses of causation outright. Halpern and Pearl's well-known account falls for similar reasons. Other accounts posit different, more "path-sensitive" structural equation analyses of causation. However, those accounts still fail when combined with their preferred counterfactualist theories of model adequacy. Consequently, I've argued, none of these accounts provide adequate definitions of causation.

Several avenues are open from here. One idea is to tie model adequacy more closely to spacetime structure. A major part of why we think that $A = 1$ doesn't cause $B = 1$ in FORKS is that $A$ and $B$ live in parts of spacetime that are, for the longest time, entirely disconnected from each other. One might hope that an approach to adequacy on which structural determination directly depends on spatiotemporal relations is sensitive to this. Given a new account of structural determination, existing model-theoretic reductions of causation should then be tested against it. I intend to take up both threads in a companion paper.

However one decides to proceed, I hope to have at least shown here that standing still isn't an option. Those who aim to provide a *definition* of causation must grapple with its consequences in worlds with causal loops. No proposal I am aware of does this adequately.

[32] THREE-WAY FORK is also a counterexample to Baumgartner's (2013) regularity-theoretic account of adequacy (given **Sufficiency**$^*$). Roughly speaking, according to Baumgartner, an adequate causal graph with variables $X$ and $Y$ contains an arrow from $X$ to $Y$ iff some value of $Y$ is part of a "minimally necessary disjunction of minimally sufficient conditions" (p. 90) for some value of $X$. (The actual definition is more involved (see pp. 90-96), but—as far as I can see—the details don't matter for our simple example.) In THREE-WAY FORK, $A = 1$ is such a disjunction for $B = 1$: $A = 1$ is both necessary and sufficient for $B = 1$, and since $A = 1$ is atomic, it is automatically both a minimal disjunction and a minimal conjunction.

Gallow's (2016) more sophisticated counterfactualist theory of adequacy arguably also struggles with causal loops. (He brackets the cyclic case in his discussion.) According to Gallow, an adequate causal model $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathcal{E})$ contains a structural equation $V := \phi_V(\mathbf{W})$ iff, for every world $w$ in the closure of the actual world's singleton set under the conditionals $\mathbf{X} = \mathbf{x} \mathrel{\square\!\!\rightarrow} ...$—one conditional for each combination of values $\mathbf{x}$ of every subset $\mathbf{X} \subseteq \mathbf{U} \cup (\mathbf{V} \backslash V)$—the (ordinary) equation $V = \phi_V(\mathbf{W})$ is true at $w$. (Additionally he imposes a maximality condition, which we'll neglect here.) Now, any adequate causal model of THREE-WAY FORK with $\mathbf{V} = \{A, B, X, Y\}$ should arguably contain the structural equations 2. But Gallow's account then requires that, if $A = 1 \wedge B = 0$, it would still be the case that $X = (A \leftrightarrow B) \cdot Y + (A \leftrightarrow \overline{B}) \cdot \overline{Y}$ and $Y = X$. But this can't be right, since the two equations are jointly logically incompatible with $A = 1 \wedge B = 0$. So Gallow's account wrongly classifies the structural equations 2 as inadequate.

# References


Al-Khalili, Jim (1999). *Black Holes, Wormholes and Time Machines*. 1st edition. Bristol, UK ; Philadelphia, PA: Taylor & Francis. ISBN: 978-0-7503-0560-0.

Albert, David Z. (2015). *After Physics*. Cambridge, Massachusetts London, England: Harvard University Press.

Baumgartner, Michael (2013). "A Regularity Theoretic Approach to Actual Causation". In: *Erkenntnis* 78.1, pp. 85–109.

Beckers, Sander and Joost Vennekens (2017). "The Transitivity and Asymmetry of Actual Causation". In: *Ergo, an Open Access Journal of Philosophy* 4.

Beebee, Helen (2004). "Causing and Nothingness". In: *Causation and Counterfactuals*. Ed. by L. A. Paul, E. J. Hall, and J. Collins. MIT Press, pp. 291–308.

Bennett, Jonathan (1984). "Counterfactuals and Temporal Direction". In: *The Philosophical Review* 93.1, pp. 57–91. (Visited on 07/25/2023).

Bernstein, Sara (2014). "Omissions as Possibilities". In: *Philosophical Studies* 167.1, pp. 1–23.

Carter, Brandon (1968). "Global Structure of the Kerr Family of Gravitational Fields". In: *Physical Review* 174.5, pp. 1559–1571.

Chalmers, David J. (2002). "Does Conceivability Entail Possibility". In: *Conceivability and Possibility*. Ed. by Tamar Szabo Gendler and John Hawthorne. Oxford University Press, pp. 145–200.

Dorr, Cian (2016). "Against Counterfactual Miracles". In: *The Philosophical Review* 125.2, pp. 241–286.

Effingham, Nikk (May 2020). *Time Travel: Probability and Impossibility*. New York: Oxford University Press.

Elga, Adam (2000). "Statistical Mechanics and the Asymmetry of Counterfactual Dependence". In: *Philosophy of Science* 68.3, pp. 313–324.

Fine, Kit (1975). "Critical Notice of Lewis, Counterfactuals". In: *Mind* 84.335, pp. 451–458.

Gallow, J. Dmitri (2016). "A Theory of Structural Determination". In: *Philosophical Studies* 173.1, pp. 159–186.

Glynn, Luke (2013). "Of Miracles and Interventions". In: *Erkenntnis* 78.1, pp. 43–64.

Gödel, Kurt (1949). "An Example of a New Type of Cosmological Solutions of Einstein's Field Equations of Gravitation". In: *Reviews of Modern Physics* 21.3, pp. 447–450.

Hall, Ned (2004). "Two Concepts of Causation". In: *Causation and Counterfactuals*. Ed. by John Collins, Ned Hall, and Laurie Paul. MIT Press, pp. 225–276.

— (2007). "Structural Equations and Causation". In: *Philosophical Studies*.

Halpern, Joseph (2000). "Axiomatizing Causal Reasoning". In: *Journal of Artificial Intelligence Research* 12, pp. 317–337.

— (2016). *Actual Causality*. MIT Press. ISBN: 978-0-262-03502-6.

Halpern, Joseph and Christopher Hitchcock (2010). "Actual Causation and the Art of Modeling". In: *Causality, Probability, and Heuristics: A Tribute to Judea Pearl*. Ed. by Halpern Joseph and Hitchcock Christopher. College Publications, pp. 383–406.

Halpern, Joseph and Judea Pearl (2005). "Causes and Explanations: A Structural-Model Approach. Part I: Causes". In: *The British Journal for the Philosophy of Science* 56.4, pp. 843–887.

Hitchcock, Christopher (2001). "The Intransitivity of Causation Revealed in Equations and Graphs". In: *The Journal of Philosophy* 98.6, pp. 273–299.

— (2007). "Prevention, Preemption, and the Principle of Sufficient Reason". In: *The Philosophical Review* 116.4, pp. 495–532.

Jäger, Jens (2021). "List and Menzies on High-Level Causation". In: *Pacific Philosophical Quarterly* 102.4, pp. 570–591.

Kajari, E. et al. (Oct. 2004). "Sagnac Effect of Gödel's Universe". In: *General Relativity and Gravitation* 36.10, pp. 2289–2316.

Lewis, David (1973a). "Causation". In: *Journal of Philosophy* 70.17, pp. 556–567.

— (1973b). *Counterfactuals*. Malden, Mass.: Blackwell.

— (1976). "The Paradoxes of Time Travel". In: *American Philosophical Quarterly* 13.2, pp. 145–152.

— (1979). "Counterfactual Dependence and Time's Arrow". In: *Noûs* 13.4, pp. 455–476.

— (1986). "Events". In: *Philosophical Papers Vol. II*. Ed. by David Lewis. Oxford University Press, pp. 241–269.

Loewer, Barry (2007). "Counterfactuals and the Second Law". In: *Causation, Physics and the Constitution of Reality: Russell?S Republic Revisited*. Ed. by Huw Price and Richard Corry. Oxford University Press.

Mandelkern, Matthew (2019). "What "Must" Adds". In: *Linguistics and Philosophy* 42.3, pp. 225–266.

Maudlin, Tim (2007). *The Metaphysics Within Physics*. Oxford University Press.

Menzies, Peter (2004). "Causal Models, Token Causation, and Processes". In: *Philosophy of Science* 71.5, pp. 820–832.

Norton, John D., Oliver Pooley, and James Read (2023). "The Hole Argument". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2023. Metaphysics Research Lab, Stanford University.

Schaffer, Jonathan (2005). "Contrastive Causation". In: *Philosophical Review* 114.3, pp. 327–358.

Stalnaker, Robert (1968). "A Theory of Conditionals". In: *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)*. Ed. by Nicholas Rescher. Blackwell, pp. 98–112.

Stockum, W. J. van (1938). "IX.—The Gravitational Field of a Distribution of Particles Rotating about an Axis of Symmetry". In: *Proceedings of the Royal Society of Edinburgh* 57, pp. 135–154.

Vihvelin, Kadri (1995). "Causes, Effects and Counterfactual Dependence". In: *Australasian Journal of Philosophy* 73.4, pp. 560–573.

Weslake, Brad (2015). "A Partial Theory of Actual Causation". In: *British Journal for the Philosophy of Science*.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in Philosophy of Science. Oxford, New York: Oxford University Press. ISBN: 978-0-19-518953-7.