# Non-Dynamical Constraints and the Reduction of Causation

Jens Jäger
New York University

Draft: November 14, 2024

**Abstract**

More than 50 years ago, David Lewis developed Hume's original idea of reducing causation to counterfactual dependence, sparking a tradition which remains influential today. This tradition has evolved: initial and relatively simple, yet counterexample-prone, theories have given way to increasingly sophisticated analyses of causation in terms of structural equation models. This paper argues that all extant counterfactual dependence analyses of causation—including those in terms of structural equation models—are false. For none correctly handle non-dynamical nomic constraints, which generate counterfactual dependence without causal dependence.

## Contents

**1 Previous Arguments Against Sufficiency** **2**

**2 Two Examples of Non-Dynamic Nomic Constraints** **6**
2.1 Gauss's Law . . . . . 6
2.2 Causal Loops . . . . . 12

**3 Troubles for Accounts of Causation** **22**
3.1 Against Lewis (1973a) and Hall (2007) . . . . . 22
3.2 Against Structural Equation Accounts . . . . . 23

**4 The Path Forward** **31**

# Introduction

AFTER now more than 50 years since Lewis's "Causation" (1973), the project of analyzing actual causation in terms of counterfactual dependence remains alive and, by many accounts, well. While the project encompasses a variety of approaches, those approaches are widely regarded to be united by a common principle. 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) ...)."

The principle is appealing. If I hadn't hit the bullseye, I wouldn't have won. If my laptop didn't have enough juice, I'd soon be sitting in front of a black screen. If you weren't reading this sentence, you wouldn't know what it says. So, one concludes, my hitting the bullseye *causes* my winning, my laptop's ample charge *causes* its continued operation, your reading the sentence *causes* your knowing what it says. Generalizing:

> **Sufficiency:** Necessarily, if $(c, e)$ is a suitable pair of actual events such that $e$ wouldn't have occurred if $c$ hadn't occurred, then $c$ causes $e$.

The restriction to "suitable" event pairs fends off otherwise easy counterexamples to the principle. For example, my hitting the bullseye doesn't *cause* my hitting the board's center—they are the same event. Yet if I hadn't hit the bullseye, I wouldn't have hit the board's center. One standard requirement on "suitability" is therefore that $c$ and $e$ be "distinct" events, meaning here that neither is a part of the other, and that neither's occurrence logically entails the other's occurrence (cf. Lewis (1973; 1979)). We'll soon see other potential suitability requirements.

There is another class of easy counterexamples to deal with. Counterfactual conditionals are notoriously context-sensitive, with Lewis (1979, p. 458) famously distinguishing between "standard" and "backtracking" contexts. Suppose that earlier today you returned Susy's book, as you promised you'd do. Given that Susy is known to take promises seriously, the following conditional seems true:

(1) If I hadn't returned the book to her 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 such retrocausal powers. Conditional (1) and conditional (2) exemplify, respectively, the "standard" and "backtracking" reading of the counterfactual conditional. For just the aforementioned reasons, it's important that the conditional in **Sufficiency** have its standard reading.

This paper argues that **Sufficiency**, and the counterfactualist tradition more broadly, face an existential threat: an underappreciated class of counterexamples, involving *non-dynamical nomic constraints*. These are constraints, imposed by physical laws, on possible configurations of dynamically unrelated events. In Section 1, I explore previous challenges to **Sufficiency** and why, despite them, **Sufficiency** has endured. In Section 2, I then provide two examples of non-dynamical nomic constraints: the first is Gauss's law of classical electrodynamics; the second involves constraints imposed by causal loops on their pasts. Section 3 argues that the failure of **Sufficiency** immediately rebuts several reductivist theories, including Lewis (1973a) and Hall (2007). While some structural equation accounts of causation don't entail **Sufficiency**, they entail a closely related principle, in terms of structural equation models. In the same section, I show that this principle fails when conjoined with the accounts' counterfactualist reduction of structural equations. In Section 4, I conclude that no extant counterfactualist reduction of causation succeeds.

# 1 Previous Arguments Against Sufficiency

**Sufficiency** has faced previous challenges. Some have argued that *omissions* aren't causes (e.g. Beebee, 2004). If that's 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. Now, 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 strengthen your notion of "suitability", entailing a restriction to pairs of *positive* events.[1]

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 (cf. Weslake (2017)). As a last resort, you could strengthen your notion of "suitability", entailing a restriction to pairs of *proportional* events.

A third challenge emerges from Lewis's own semantics for counterfactuals in deterministic worlds. It turns out that his original choice of labels, "standard" vs. "backtracking", is misleading: even on his own, miracles-based, semantics, the "standard" resolution of the counterfactual conditional involves backtracking, i.e. counterfactual past differences. For the closest possible antecedent worlds usually includes *ramping periods*. Specifically, according to Lewis, counterfactual antecedents are preferentially brought about by small miracles (cf. (Lewis, 1979)), and small miracles need time to snowball into big change. So, where antecedents dictate big differences to actuality, they must include a significant delay between miracle and antecedent event—a delay 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 miss the ball. Instead, some macroscopic change would have occurred—perhaps I would have noticed the ball only later, or you threw it a littler harder, or a gust of wind deflected the ball outside of my reach. According to Lewis, any of these changes would be brought about by a small miracle—e.g. changes in neuronal firing patterns, or microscopic meteorological changes—which subsequently needs time to effect the big change. Again, this raises the specter of retrocausation: my failure to catch the ball certainly causes neither my earlier neurons' firing nor the earlier atmospherical state.

Lewis (1979) offers a response. According to **Sufficiency**, $c$ causes $e$ if $\neg O(c) \mathrel{\Box\!\!\rightarrow} \neg O(e)$ (where $\ulcorner O(x) \urcorner$ denotes the proposition that $x$ occurs and " $\Box\!\!\rightarrow$ " is the counterfactual conditional). On Lewis's semantics, this requires that $\neg O(e)$ in *all* closest worlds where

[1] This restriction could also be subsumed under a ban on "overly disjunctive" events, as e.g. (Lewis, 1986b, p. D) discusses; see below.

$\neg O(c)$. But an event's mere *non*-occurrence typically leaves much undetermined. Lewis's hope is that this includes the ramping period's content:

> "[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, however, 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 given event hadn't occurred]."[2] (Lewis, 1979, p.463)

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. The hope is that, for any actual positive event $e^*$ preceding $c$, there is a way of filling out the ramping period in which $e^*$ still occurs. Then $\neg O(c) \mathrel{\square\!\!\rightarrow} \neg O(e^*)$ is false for any such $e^*$.[3]

Vihvelin (1995) identifies two kinds of threats to this response. The first arises when the antecedent event is an *omission*—an omission's negation typically entails definite ramping period. But this is no threat to the weakening of **Sufficiency** to positive events. The second threat stems from overly detailed past events. Let $c$ be *my catching the ball at* $t$, for some time $t$, and let $e$ be the totality of all events during some open time interval bounded, to the future, by $t$. Given that the laws are deterministic, necessarily $e$ doesn't occur if $c$ doesn't, and so $\neg O(c) \mathrel{\square\!\!\rightarrow} \neg O(e)$. But my catching the ball doesn't *cause* $e$.

One possible response follows Lewis (1986a) in banning overly "fragile" events. Those are events with extremely detailed essences—intuitively, events which could have very easily failed to occur. Lewis's justification for the ban is that our standard way of denoting event propositions, "standard nominalizations", isn't nearly detailed enough to pick out these events. Inspired by Lewis's move, one might strengthen "suitability" yet again, additionally requiring that $c$ and $e$ not be *overly fragile* events.

Lewis's responses are unlikely to convince everyone: it has not been conclusively shown that *all* non-occurrences of robust, positive events entail eclectic swaths of equally-

[2]The original quote ends with "...if the past were different". This is stronger and less plausible than what Lewis needs. For example, in some contexts "...if the past were different" might be coextensive with ⌜...if $p$ hadn't occurred⌝ where ⌜$p$⌝ picks out some omission. But we'd generally expect the non-occurrence of an omission—i.e., the occurrence of a positive event—to fix the past rather definitely: for example, if it hadn't been the case that I didn't go to the fridge now—i.e., if I *had* gone to the fridge now—I would have gotten up shortly before. I suspect my substitution better captures what Lewis actually has in mind.

[3]This strategy hinges on further particulars of Lewis's semantics. On a semantics like Stalnaker's (1968), in which any utterance of a conditional entails a unique closest antecedent world, there's a unique closest-possible way of filling out the ramping period. In this case some positive event $e^*$ preceding $c$ may well satisfy $\neg O(c) \mathrel{\square\!\!\rightarrow} \neg O(e^*)$. In response, one could hold that, even though the conditional is true, it's not *determinately* true. A weakened version of **Sufficiency** which requires *determinate* truth of $\neg O(c) \mathrel{\square\!\!\rightarrow} \neg O(e^*)$ would then survive.

close ramping periods. And as for the distinction between the robust and the fragile, Lewis provides no independent characterization. Some counterfactual reductivists have therefore proposed an alternative: retreating to a variant of **Sufficiency** where the counterfactual conditional doesn't require ramping periods. I'll explain this alternative strategy in the following footnote.[4]

Now, a *fourth* challenge results from *rejecting* counterfactual miracles, in favor of a view on which counterfactual worlds have the same laws, but different pasts all the way back (cf. Bennett, 1984; Loewer, 2007; Albert, 2015; Dorr, 2016). The view holds that, assuming I don't get up from my chair, then if I did get up, the world would be different arbitrarily far back in time. Notably, in a world like ours, with continuous laws, the past will only have to be *microscopically* different, until very close to the cause's time (cf. Dorr (2016)). But surely, distant retrocausation—even if its effects are microscopic—isn't as easy as getting up from my chair. So **Sufficiency** is wrong. In my view, the cleanest way for the **Sufficiency** advocate to resist this challenge is to stick with Lewis's miracles-based semantics (or variants of it which avoid ramping periods). Indeed, miracles-based semantics are standard in the counterfactual dependence tradition (cf. Lewis (1979), Hitchcock (2001), Hall (2007), and Glynn (2013)).

[4] Glynn's (2013) account is an example of this. It's a combination of two ideas. First, stipulate a technical meaning of the counterfactual conditional—let's denote it $\blacksquare\!\!\rightarrow$ —according to which, when $A$ is a proposition purely about particular matters of fact at the instant (or brief interval) $t$, "$A \blacksquare\!\!\rightarrow B$" is evaluated using only miracles *at* $t$. That is, $A \blacksquare\!\!\rightarrow B$ is true iff $B$ is true at all worlds closest among those where (i) $A$ is true, (ii) no miracles occur outside of $t$, (iii) everything prior to $t$ is as it actually is, and (iv) everything after $t$ evolves according to the actual laws of nature. Second, say that $c$ *causes* $e$ if there is some truth $T$ solely about $t$ such that $\neg O(c)$ and $T$ are metaphysically compossible and $\neg O(c) \wedge T \blacksquare\!\!\rightarrow \neg O(e)$, where $\blacksquare\!\!\rightarrow$ is evaluated in the technical way above. Intuitively, the role of $T$ is to suppress any unwanted effects which the $\neg O(c)$-realizing miracle would otherwise bring about—*viz.* consequences which affect $e$ via causal routes bypassing $c$. Generally, $T$ requires a highly complex miracle. For example, my bus is stuck in traffic, and my being on the bus right now ($c$) causes my being late to the meeting ($e$). Moreover, if I wasn't on the bus right now, I'd be on my bike ($\neg O(c)$), speeding through grid-locked traffic, and arriving on time. On the technical meaning of the counterfactual conditional introduced above, if I was on my bike right now, this would be because a miracle had instantly teleported me from the bus onto the bike. But such a miracle would have all sorts of undesired byproducts, affecting my arrival time independently of my newly acquired ability to cycle there. For concreteness, suppose the miracle would leave me extremely startled—so much so that I'd crash, thus not arriving on time. Still, we want to say that my being on the bus ($c$) causes my being late ($e$). Glynn's account achieves this as follows. Some $T$ entail that I'm calm and not disoriented, thus negating the unintended consequence. Moreover, Glynn makes it plausible that we can find a $T$ to do this for all unintended consequences (including consequences that arise from the miracles needed to bring about the various suppressors). If so, then if $\neg O(c) \wedge T$, then 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) without the need for counterfactual ramping periods. The idea would then be to change **Sufficiency** as follows:

> **Sophisticated Sufficiency:** Necessarily, if $(c, e)$ is a suitable pair of actual events such that, for some truth $T$, $\neg O(c) \wedge T \blacksquare\!\!\rightarrow \neg O(e)$, then $c$ causes $e$.

Importantly, however, **Sophisticated Sufficiency** still falls to my counterexample, as I'll explain in fn. 24.

(Alternatively, the **Sufficiency** advocate could strengthen her notion of "suitability", restricting it to pairs whose second element (the putative effect) is a *macroscopic* event. The downside is that this is vulnerable to cases with non-continuous laws. One could imagine a universe with discretized Newtonian gravity, for example, where any difference in the gravitational force between two bodies requires macroscopically different masses. In such a universe, if the force between two sufficiently isolated bodies had been different at $t$, the two bodies would have to have had macroscopically different masses already before $t$. But the force at $t$ doesn't *cause* them to have the masses they actually have before $t$. In response, the **Sufficiency** advocate might argue that the case for non-miracles-based counterfactual semantics is weaker in worlds with discrete laws, because one of their motivations—that only *microscopic* counterfactual adjustments to the past are required to accommodate the antecedent—no longer applies. Be that as it may; since miracle-based semantics are standard among counterfactualist reductions of causation anyway, I'll just stick with it.)

**Sufficiency** emerges weakened but still makes interesting and substantive predictions: the canonical examples of causation, which also motivate counterfactual reductions, tend to involve positive, proportional, and not overly fragile 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 in its weakened form, **Sufficiency** is false. This is because, in the presence of *non-dynamical nomic constraints*, even *standard* counterfactual dependence fails to track causal dependence.

# 2 Two Examples of Non-Dynamic Nomic Constraints

## 2.1 Gauss's Law

Consider a universe with Maxwell's laws of electrodynamics, containing a single stationary electron at spatial location $x$;[5,6] otherwise the universe is completely empty of particles. Call this world GAUSS. Pick some arbitrary time $t$. The following conditional seems clearly true:

[5]In the following, any references to spatial location, time, geometrical properties like sphericalness and isotropy, and other Lorentz non-invariant properties, are relative to a fixed reference frame co-moving with the electron.

[6]To avoid discontinuity in the resulting electric field, assume that the electron has some very small but non-zero spatial extent and that the electric field density falls off continuously near $x$.

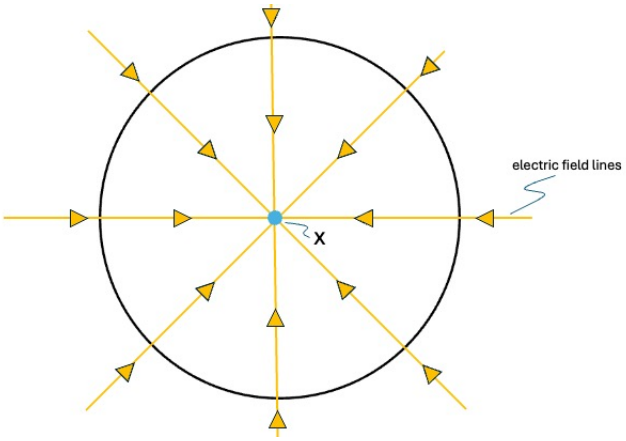



Figure 1: A 2D sketch of (three-dimensional) GAUSS

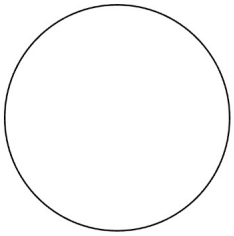
**(C)** If the electron didn't exist at location $x$ at $t$, the universe would be completely empty of particles at $t$.

But Maxwell's laws entail Gauss's law, which relates the electric field at the boundary of any spatial volume to the amount of electric charge contained in the volume. Intuitively, it implies that when there is zero net charge inside the volume, the electric field lines at the volume's boundary which point *into* the volume are perfectly balanced out by those pointing *out of* the volume. Formally, this balance is known as the *electric flux* through the boundary.[7] Meanwhile, if there is a non-zero net charge inside the volume, then (depending on whether the charge is positive or negative) more field lines point out of the volume than point inward, or *vice versa*—i.e., the electric flux through the boundary is either positive or negative.

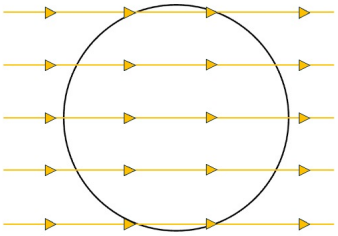
Now pick an arbitrarily large (hollow) sphere $S$ centered on the electron. Since $S$ encloses a net negative electric charge, there must be a non-zero electric flux through $S$. (In particular, since the electron is stationary, the electric field density is rotationally symmetric around $x$. See fig. 1 for a 2D sketch.) Now, according to **(C)**, if it wasn't for the electron at location $x$ at $t$, there would be no electric charge at $t$. Maxwell's equations would remain true at any time $t^+ > t$—even fans of a miracle-based semantics should agree to as much. Specifically, choose $t^+$ such that $S$ at $t^+$ is outside of $(t, x)$'s future light-cone.[8] Since Maxwell's equations entail charge conservation, it follows that, if it wasn't for the electron at location $x$ at $t$, there would be no electric charge at $t^+$. But Gauss's law—part

[7] More precisely, the *electric flux* through a (spatial) surface is the integral, over the surface, of the scalar product of electric field and the boundary's outward-facing normal. Gauss's law states that the electric flux through a volume's boundary is proportional to the net amount of charge enclosed in the volume.

[8] The *future/past light-cone* of a spacetime point $p$ is the largest spacetime region such that every point in it can be reached from $p$ via a future-directed/past-directed causal curve. (Some authors call these regions instead $p$'s *causal future/past*.) A curve is *causal* if its tangent vector is everywhere either time-like or null.



(a) A charge-free solution to Maxwell's laws (2D sketch).



(b) Another charge-free solution to Maxwell's laws (2D sketch)

of Maxwell's equations—would still be true at $t^+$, and so the electric flux through $S$ at $t^+$ would be zero and hence different from what it actually is. Hence the distribution of electric field lines over $S$ at $t^+$ would be different from what it actually is. See figs. 2a and 2b for sketches of two solutions with zero electric flux through $S$. (We needn't settle which zero-flux solution is among the closest counterfactual worlds.)

So, where $c_G$ is the electron's being located at $x$ at $t$, and $e_G$ is the electric flux through $S$ at $t^+$'s being non-zero, we have

$$\neg O(c_G) \mathrel{\Box\!\!\rightarrow} \neg O(e_G).$$

Hence, by **Sufficiency**, $c_G$ causes $e_G$. But Maxwellian electrodynamics is a *local* theory—electromagnetic influence by subluminal particles travels at most with the vacuum speed of light, and so the electron at $x$ at $t$ doesn't causally influence the electric flux through $S$ at $t^+$. In other words, $c_G$ doesn't cause $e_G$. Hence **Sufficiency** is false.

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

$1_G$. If **Sufficiency** is true, GAUSS is possible, $(c_G, e_G)$ is a suitable pair of events at GAUSS, and $\neg O(c_G) \mathrel{\Box\!\!\rightarrow} \neg O(e_G)$ in GAUSS, then $c_G$ causes $e_G$ in GAUSS.

$2_G$. $(c_G, e_G)$ is a suitable pair of events at GAUSS.

$3_G$. **Possibility**$_G$: GAUSS is possible.

$4_G$. **Dependence**$_G$: $\neg O(c_G) \mathrel{\Box\!\!\rightarrow} \neg O(e_G)$ in GAUSS.

$5_G$. **Non-Causation**$_G$: $c_G$ does *not* cause $e_G$ in GAUSS.

$\therefore$ **Sufficiency** is false.

Premise $1_G$ follows from the fact that a necessary truth is true in all possible worlds. What about Premise $2_G$? We've seen several demands the **Sufficiency** advocate might place on

"suitability". $c_G$ and $e_G$ are certainly *distinct*, occurring as they do in separate spacetime regions. They are also *proportional* to each other, both being simple configurations of physically fundamental properties. Likewise, neither event is overly "fragile": both the electron's being located at $x$ at $t$ and the electric flux through $S$ at $t^+$'s being negative are specifiable by ordinary nominalizations (as we just did), and so not "fragile" in Lewis's intended sense. Moreover, we could have made the same argument with more approximate spatial and temporal locations. Approximate particle locations and electric fluxes, measurable in a lab, are prime examples of physical facts we'd like to *causally explain*. So any theory of causation which excluded them would be a non-starter.

I'll now turn to premises $3_G - 5_G$.

### 2.1.1 Defending *Possibility*$_G$

Maxwell's laws of electrodynamics are clearly metaphysically possible, and their content metaphysically compossible with the existence of a single electron and a static electric field. To conclude that GAUSS is metaphysically possible, all we need in addition is that the existence of a single electron and a static electric field is metaphysically compossible with Maxwell's laws of electrodynamics being *laws.* Primitivists about laws won't struggle with this point.

But nomic reductivists might complain. A Humean best-system account of lawhood—arguably the most popular nomic reductivist approach—will hold that Maxwell's laws, including Gauss's law, are too complicated. Systems explicating the particle's actual position and accompanying field configuration in GAUSS purchase more strength for less complexity.

But there's an easy counter: consider GAUSS$^*$, a universe just like GAUSS except with complex electrodynamical systems far, far away from our particle. In GAUSS$^*$, Gauss's law is still true. The only difference is that a nomic reductivist will now also agree that it is a *law*. The moderate complexity of Maxwell's equations purchases significant strength, by constraining the available possibilities for far away systems, achieving a better strength-simplicity tradeoff than any system that didn't include them.

### 2.1.2 Defending *Dependence*$_G$

We've already laid out the argument for **Dependence**$_G$: if the electron didn't exist at $x$ at $t$, no charge would exist from $t$ onward (in particular, it wouldn't spontaneously reappear); yet Gauss's law would still hold and thus the electric field over $S$ would have to be different at all times after $t$.

The only remaining way to preserve *local* counterfactual dependence between electron and field would be to admit that Gauss's law *never* holds after $t$. For as long as it held at *some* $t^* > t$, there would be a spatial location $x^*$ such that $(t^*, x^*)$ is outside of $(t, x)$'s future light-cone yet the electric field would be different from actuality at $(t^*, x^*)$. To avoid **Dependence**$_{\text{G}}$, therefore, we must grant that the closest $\neg O(c)$-worlds contain a temporally infinite miracle, breaking the actual laws *at t and forever after*.

Now, you might reply, **Dependence**$_{\text{G}}$ *also* requires an infinite counterfactual miracle—a spatial one: to preserve Gauss's law in absence of charge, the electric field must be different from actuality at every spatial radius from $x$ at $t$. So, the electron's counterfactual absence requires an infinitely large miracle either way.

But there are three reasons why opponents of **Dependence**$_{\text{G}}$ are still worse off. First, the literature on counterfactualist reduction literature often explicitly rejects temporally extended miracles. For example, Glynn's (2013) account confines miracles to a single moment (or brief interval) in time—the time of occurrence of the event in the counterfactual's antecedent. It's not obvious how to adjust his "late-miracle" account to also allow miracles outside of that time.

Second, the closest possible worlds suspending Gauss's law aren't what one might initially expect. In conversation, the most common suggestion has been that the electric field would be 0 at $(t, x)$, with the sphere of vanishing electric field expanding at light speed thereafter. But there is little basis for thinking this. Suppose we follow the suggestion and remove the electron at $(t, x)$, set the electric field to 0 at $(t, x)$, smooth it out near $x$, and leave everything else unchanged. Applying Maxwell's equations, *sans* Gauss's law, to this state simply yields the static solution: the electric field remains unchanged forever after $t$.[9] There is no basis in Maxwell's laws for the suggestion of a counterfactually expanding sphere of vanishing electric field.

Third, that the closest counterfactual worlds which suspend Gauss's law are static is independently troubling for my opponent. It would follow for her that, outside of a tiny neighborhood around $x$, the electric field in $(t, x)$'s future light-cone doesn't counterfactually depend on the electron's presence at $(t, x)$. Now, causation doesn't necessitate counterfactual dependence. But it does *imply* it *if* the causal structure is simple enough: specifically, if there is no possibility for preemption or overdetermination. (Modern counterfactual dependence analyses concur.) But in GAUSS we have neither: the electron doesn't preempt a back-up cause of the future field, nor does its presence causally overdetermine

[9]To see this, note that the magnetic field vanishes everywhere at $t$ in GAUSS, and thus still vanishes in the counterfactual world. Since the electric field $\mathbf{E}$ at $t$ still only has a radial component, its curl vanishes. Hence the magnetic field $\mathbf{B}$ remains at 0 (via Faraday's law, $\text{curl}(\mathbf{E}) = -\frac{\partial \mathbf{B}}{\partial t}$), and so the electric field doesn't change either (via Ampere's law for vanishing electric current, $\text{curl}(\mathbf{B}) = \frac{\partial \mathbf{E}}{\partial t}$).

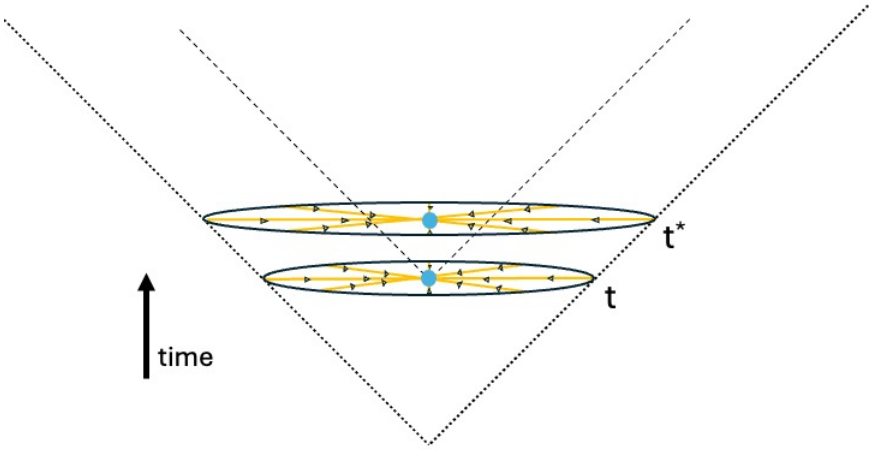



Figure 2: Sketch of GAUSS$^\dagger$, past-ward bounded (one spatial dimension suppressed)

it. As a result, *if* the given static universe was the closest possible $\neg O(c)$-world, one would have to conclude that the electron's presence at $x$ at $t$ is, besides a tiny neighborhood around $x$, entirely causally inert. But that's absurd.

In short, then, preserving local counterfactual dependence by positing a temporally infinite counterfactual miracle faces serious obstacles. The initially intuitive option, of preserving Gauss's law and relinquishing local counterfactual dependence, remains attractive.

Those still on the fence may consider a modified version of the case, GAUSS$^\dagger$: a past-ward bounded spacetime, consisting exactly of the interior of a single point's future light-cone.[10] Consider two three-dimensional bounded cross sections, $t$ and $t^*$, of this light-cone; cf. fig. 2. In GAUSS$^\dagger$, a spatially *finite* miracle would restore Gauss's law if the electron vanished at $t$—for example, it might set the electric field to zero everywhere on $t$. Meanwhile, proponents of local counterfactual dependence, leaving the electric field outside of $(t, x)$'s future light-cone unchanged, would still require breaking Gauss's law forever after $t$.

A second objection targets the conditional **(C)**—"If the electron didn't exist at location $x$ at $t$, the universe would be completely empty of particles at $t$"—a crucial part of our argument for **Dependence$_G$**. I've asserted that **(C)** is clearly true. But here is a worry. As previously mentioned, Maxwell's equations also imply the conservation of electric charge: the net outflow of electric charge through the surface of a volume must equal the net decrease of electric charge enclosed by the volume. Doesn't that mean that the

[10] More properly: the manifold is *isometric*, rather than numerically identical, to the future light-cone of a single point—for there is no point of which the manifold is a future light-cone.

deletion of the electric charge should be compensated somewhere? Note that, even if it did, then wherever those compensating charges appeared, say at some radius $r$ from $x$, our argument would still go through for a sphere $S$ of radius $< r$. But, regardless, the answer is "no" anyway: the miracle which deletes the electron at $t$ also breaks the conservation law at $t$ (with the law then resuming to hold at all times afterward). This is entirely in line with our ordinary counterfactual reasoning with conservation laws: actually, the electron has zero momentum at $t$; on the standard reading, we won't be at all inclined to say that, if the electron had non-zero momentum at $t$, there would be another particle present at $t$ with equal and opposite momentum. Instead, either a small miracle would have jerked the particle, or (on views which hold the laws fixed) the particle would have had the non-zero momentum all along. So the objection against **(C)** is simply mistaken: disappearing the electron at $t$ also breaks the conservation of electric charge at that instant.

### 2.1.3 Defending *Non-Causation*$_{\mathrm{G}}$

**Non-Causation**, as we said, is simply a consequence of the universally accepted claim that Maxwell's theory is *local*, involving no action at a distance. To see how this claim is compatible with Gauss's law, note that the electric field configuration over $S$ at $t^+$ is completely determined, via Maxwell's equations, by any cross-section of its past light-cone.[11] But $(t, x)$ isn't part of the past light-cone of $S$ at $t$, and so the particle's presence at $(t, x)$ is dynamically irrelevant to the electric field configuration over $S$ at $t$.

To nonetheless posit a causal influence of the particle's presence at $(t, x)$ on the electric field over $S$ at $t$ would thus be to posit an egregious form of causal overdetermination—one which introduces a particularly theoretically costly, since non-local, interaction—all while yielding zero explanatory benefit.

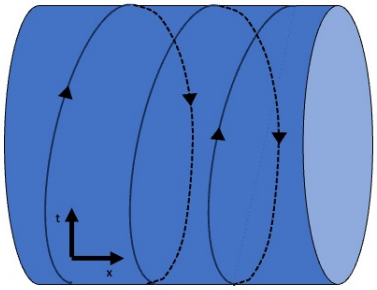
## 2.2 Causal Loops

So much for Gauss's law. Our second example involves non-dynamical nomic constraints of a different character, not due to an explicit synchronic law, but instead due to the world's global geometric structure.

*Causal loops* are hypothetical situations where a causal influence travels back in time and interacts with its source. Formally, they are chains of events $(e_1, ..., e_n)$ such that, for all $i = 1, ..., n-1$, $e_i$ causes $e_{i+1}$, and $e_n$ causes $e_1$. 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)). But this neglect

[11] The past/future light-cone of an extended region is the union of all past/future light-cones of its points.

is in tension with the goal of providing an *analysis* of causation. While causal loops are peculiar, causal judgments about them aren't mere "spoils for the victor". Our intuitions about causation and counterfactuals on causal loops are lucid enough that, provided such scenarios are metaphysically possible, they ought to be accommodated by any serious reductive account of causation. 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.[12]

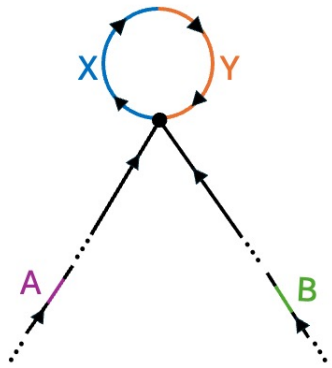



Suppose the laws say that the particle moves uniformly, with (four-)velocity $\mathbf{u}$. It seems obvious—or at least as obvious as in the acyclic case—that the particle's being located at $\mathbf{x}$ *causes* its being located at $\mathbf{x}^+ := \mathbf{x} + \mathbf{u} \cdot \Delta\tau$ after (proper)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

Now, it's unsurprising that causal loops impose constraints on time-slices that *intersect* them. To illustrate: a particle's trajectory generically intersects all time-slices of our cylinder at infinitely many different locations. Hence the existence of a particle at some location $\mathbf{y}$ on a time-slice $t$ generically entails its existence at infinitely many more locations in $t$. But these constraints don't obviously pose problems for **Sufficiency**: after all, the infinitely many particle locations *are* connected as causes and effects.

We're therefore interested in a different kind of case: one where causal loops impose constraints on their *pasts*. Consider LOOP, a one-dimensional spacetime with the following topology:

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




The structure can be constructed by taking a one-dimensional oriented spacetime with an inverse "three-way fork" and "identifying" the upper boundary (end point) of $X$ with the lower boundary (starting point) of one of the incoming spokes, $Y$. 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 three lines merge into one (indicated by the black dot in the above figure).[13] Boring points are all remaining 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 field on an odd number of incoming lines has value 1, then the value at the fork point is 1; otherwise it is 0. Let's introduce for each region a binary variable of the same name, whose value represent the fields value taken at that region. We can then write the dynamics compactly as follows (where $\overline{\phi}$ abbreviates $(1-\phi)$):

$$X = Y \cdot (AB + \overline{A}\overline{B}) + \overline{Y} \cdot (\overline{A}B + A\overline{B}) \tag{1}$$

$$Y = X. \tag{2}$$

Suppose that actually $A = B = 1$ and $X = Y = 0$. Call this world LOOP.

The dynamics has exactly four solutions:

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

| $A$ | $B$ | $X$ | $Y$ |
|---|---|---|---|
| 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.

Now suppose that, actually, $A = B = X = Y = 1$, and let $c_{\mathrm{L}}$ be the event of $A$'s taking value 1 and $e_{\mathrm{L}}$ the event of $B$'s taking value 1. I claim that we have the following:

$$\neg O(c_{\mathrm{L}}) \mathrel{\Box\!\!\rightarrow} \neg O(e_{\mathrm{L}}),$$

or, written in terms of variables,

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

Hence, **Sufficiency** implies that $c_{\mathrm{L}}$ is a cause of $e_{\mathrm{L}}$. But, I say, $c_{\mathrm{L}}$ isn't a cause of $e_{\mathrm{L}}$. So Sufficiency is false.

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

$1_{\mathrm{L}}$. If **Sufficiency** is true, LOOP is possible, $(c_{\mathrm{L}}, e_{\mathrm{L}})$ is a suitable pair of events at LOOP, and if $\neg O(c_{\mathrm{L}}) \mathrel{\Box\!\!\rightarrow} \neg O(e_{\mathrm{L}})$ in LOOP, then $c_{\mathrm{L}}$ causes $e_{\mathrm{L}}$ in LOOP.

$2_{\mathrm{L}}$. $(c_{\mathrm{L}}, e_{\mathrm{L}})$ is a suitable pair of events at LOOP.

$3_{\mathrm{L}}$. **Possibility**$_{\mathrm{L}}$: LOOP is possible.

$4_{\mathrm{L}}$. **Dependence**$_{\mathrm{L}}$: $\neg O(c_{\mathrm{L}}) \mathrel{\Box\!\!\rightarrow} \neg O(e_{\mathrm{L}})$ in LOOP.

$5_{\mathrm{L}}$. **Non-Causation**$_{\mathrm{L}}$: $c_{\mathrm{L}}$ does *not* cause $e_{\mathrm{L}}$ in LOOP.

$\therefore$ **Sufficiency** is false.

Premise $1_{\mathrm{L}}$ again follows from the fact that a necessary truth is true in all possible worlds. Premise $2_{\mathrm{L}}$ follows because $c_{\mathrm{L}}$ and $e_{\mathrm{L}}$ are clearly distinct, positive, proportional to each other, and not too fragile; so they constitute a suitable event pair. The arguments for the remaining premises differ from those in the previous section.

### 2.2.1 Defending *Possibility*$_{\text{L}}$

**Possibility**$_{\text{L}}$ asserts that LOOP is (metaphysically) possible. One might resist this idea by denying that causal loops are metaphysically possible. But there are strong reasons to think that they *are* metaphysically possible, and little reason to think they aren't.

I take well-understood mathematical models of spacetimes to be powerful (if defeasible) guides to metaphysical possibility. One possible way to cash this out is in terms of positive conceivability. Roughly, where negative conceivability involves a mere imagining of the absence of a contradiction, positive conceivability involves, additionally, an imagining of a "positive picture" of a situation (Chalmers, 2002). The positive conceivability-possibility link then states that, if a situation is positively conceivable, it is metaphysically possible. This view is immune to counterexamples afflicting a naive link from conceivability *simpliciter* to possibility: for example, unprovable mathematical truths (such as, perhaps, Goldbach's conjecture) are metaphysically necessary, and their falsity arguably negatively, but not positively, conceivable. Situations that are represented by fully interpreted mathematical models—like LOOP—are paradigm cases of positive conceivability: the models present a precise and detailed positive picture.[14] So we should think that those situations are metaphysically possible.

In addition to considering conceivability in the abstract, note that the Einstein equations themselves have solutions with closed causal curves. In Kurt Gödel's (1949) famous example, suitably accelerated material bodies can travel along closed time-like curves. Other well-known 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).[15] Usually, we are suspect of attempts to dismiss, on *a priori* grounds, a substantial part of the scientific literature as concerned with the metaphysically impossible.

Of course, this charge of philosophical overreach could be countered if there were strong positive arguments for the impossibility of causal loops. One of the more influential worries stems from cognates of the grandfather paradox. Let *autoinfanticide* be the act

[14] "Fully interpreted" is doing work here. Consider the debate about haecceitism in the metaphysics of spacetime: do swaps of mathematical points correspond to possible swaps of spacetime points? (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.) This question arises because it's not fully settled what swaps of mathematical points are supposed to represent. But the representational aspects of LOOP which matter to us 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 is supposed to represent a closed time-like path in spacetime. These sorts of facts are all we need.

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

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 a contradiction is possibly true: it's possibly both the case and not the case that you can commit autoinfanticide. 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 may say that a speaker $S$ can $\phi$ in context $C$ iff $S$'s $\phi$-ing is metaphysically compossible with $C$. There are then plenty of contexts in which I *can* kill the baby in the crate: namely most contexts which omit that the baby is me. They witness the fact that "I have what it takes", but don't imply that I can commit *autoinfanticide*: they merely show that I can kill someone who's not me. This resolves the apparent contradiction.

Another family of objections concerns the "bootstrapping" aspect of causal loops. One version of this worry complains 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. For other versions of the no-bootstrapping worry, and rebuttals against them, see also Effingham (2020, Ch. 5.2.2).

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 LOOP. So LOOP is metaphysically possible.

### 2.2.2 Defending *Dependence*$_{\mathrm{L}}$

Say that a spacetime point is "upstream" from a spacetime region iff there is a future-directed causal curve from the point into the region. Recall that $\neg O(c_{\mathrm{L}})$ and $\neg O(e_{\mathrm{L}})$ are the propositions that $A = 0$ and $B = 0$. **Dependence**$_{\mathrm{L}}$ follows from the conjunction of two counterfactual conditionals:

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

(b) If $A = 0$, then the dynamics everywhere not upstream of $A$ would have been the

same.

By the identity rule (i.e., $\vdash \alpha \mathrel{\square\!\!\rightarrow} \alpha$) and agglomeration,[16] (a) and (b) jointly entail

(c) If $A = 0$, 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 $A = 0$, 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 it's entailed by any account of counterfactuals which holds the laws fixed (cf. Section 1).

What about miracles-based semantics—the leading semantics among counterfactualist accounts of causation? On Lewis's (1979) account, as well as alternatives like Glynn (2013), miracles are confined to *no later* than the time of the antecedent.[17] Breaking with this view, by allowing miracles *downstream* from $A$ seems unattractive: any principled reason for placing a miracle downstream from $A$ in LOOP would presumably also carry over to placing a miracle (say) in the future of Nixon's pressing the button (cf. Fine (1975)), thus defeating counterfactualist reductions of causation in ordinary cases. Moreover, miracles *prior* to $A$ (on the $A$-branch) are obviously useless for keeping $B$ fixed at 1. A miracle prior to $B$ (on the $B$-branch) doesn't help either: setting the field prior to $B$ to 1, in a world where $A$ is 0, leads to contradiction.

The only remaining option is a miracle occurring *after* $B$, on the interior of $B$'s branch, changing the field value back to 0. The resulting world would witness $A = 0 \wedge B = 1$. 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**. More importantly, such a miracle ultimately doesn't solve the problem, which simply recurs for intervals *between* the post-$B$ miracle and $Y$: let $B'$ be such an interval. 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-$B'$. So this strategy won't work.

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

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

[17] Now, Elga (2000) has shown (in my view 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, and instead just grant that there's *some*—as yet unspecified—reductive semantics which produces the desired asymmetry of miracles. There is no analogous problem with Glynn's (2013) semantics, which confines miracles to exactly the time of the antecedent.

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 $A$. In such a world, $A = 0 \wedge B = 1$ would be compatible with the laws.

Unfortunately, strategy (i) again simply pushes the problem back: we can restate the problem for any interval $B'$ *between* the cut and $Y$. Strategy (ii), meanwhile, assumes that worlds with topology changes downstream from $A$ are among the closest possible $A = 0$-worlds. But topology changes function much like miracles; if they were counterfactually cheap, we'd be back at the Nixon problem, threatening counterfactualist analyses of causation in ordinary cases: Nixon's button press wouldn't have led to nuclear war because the signal in Nixon's cable would have 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 three-way fork point into a two-way fork point. But some possible dynamics for the remaining two-way fork point forbid $A = 0$: for example, $X = AY + \overline{AY}$ together with $Y = X$ entails that $A = 1$. We can simply amend LOOP by specifying this dynamics for two-way fork points. (For the Humean, we do this by adding appendages to the spacetime, containing many two-way fork points exhibiting this dynamics.) In this case, no $A = 0$-world in which the $B$-branch is severed from the rest of spacetime is closest to actuality.

I conclude that Premises (a) and (b) both stand. Now, in the introduction I emphasized the importance of *standard* readings for **Sufficiency**. Does the present argument somehow rely on interpreting the premises according to a backtracking reading? No: Premises (a) and (b) explicitly concern only the part of LOOP not to the past of $A$. Moreover, intuitively, the conditionals in (a) and (b) don't require any special accommodation on the part of the listener—there's no sense of the context shift so characteristic of backtracking readings. Unlike, for example, with conditional (2)—where the switch to a backtracking reading is intuitively noticeable—in the case of conditionals (a) and (b) we remain in our standard mode of reasoning.

To remove any ambiguity about this, we can also modify LOOP: first add additional temporal structure to LOOP, 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 LOOP. 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 $B = 1$: in particular, the Weak Intrinsicness argument from the last section carries over without issue. But now it's entirely unambiguous that the counterfactuals (a) and (b), which all remain true, are non-backtracking. For there is simply nowhere to backtrack to—there's no universe prior to $A$.[18]

### 2.2.3 Defending *Non-Causation*$_{\mathrm{L}}$

$A$ and $B$ occur on initially separate branches of spacetime. Up until they occur, and for a long time afterwards,[19] there is no spatiotemporal connection whatsoever between the two lines. Up until the very far future, when the two lines finally merge, there is no telling that they ever do so. But what happens (say) a billion years out shouldn't matter for whether $A = 1$ causes $B = 1$ *now*.

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 Intrinsicness:** Let $w$ and $w'$ be worlds with the same laws and with identical histories up to (and including / excluding) time $t$.[20] Then, for any

[18] 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" improves felicity; as, for example, in the following case: "if I hadn't returned the book today, it would *have to have been* the case that Susy and I agreed on a later date to begin with." But performing the replacements on (a) and (b) arguably doesn't improve felicity:

(a*) If $A = 0$, then the spacetime structure everywhere not upstream of $A$ would have to have been the same.

(b*) If $A = 0$, then the dynamics everywhere not upstream of $A$ would have to have been the same.

In any case, I don't think the substitution test reliably distinguishes standard from backtracking, or more generally *causal* from *non-causal*, readings anyway. Instead, I think "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. To give an example where the epistemic modal is embedded in a subjunctive conditional: suppose you're skeptical that your child was home yesterday (and indeed she wasn't home). You know that your neighbor set off loud fireworks at night. Yet, when you ask your child, she claims not to have heard any loud noises. So you insist: "If you had been home yesterday, you *would have to have* heard some loud bangs." This counterfactual embeds "have to", yet is causal: not being home caused her not hearing the bang. So the substitution test plausibly doesn't separate causal from non-causal counterfactuals. (Many thanks to Richard Roth for helpful discussion here.)

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

[20] "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 will just be slightly more cumbersome to state.

events $c$ and $e$ in $w$ occurring up to (and including / excluding) time $t$, if it's true at $w$ that $c$ causes $e$, then it's true in $w'$ that $c$ causes $e$.

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

> **Intrinsicness (Hall):** Let world $w$ contain $S$, a "structure of events that consists of $e$, together with all of its causes [in $S$] back to some arbitrary earlier time $t$" (ibid., p. 239). Let $c$ be some cause of $e$ in $w$. Then, if $w'$ has the same laws as $w$ and contains $S$,[21] then $c$ causes $e$ in $w'$.

Hall's Intrinsicness 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 an identical 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 in the alternative world. So Hall's Intrinsicness principle would predict, wrongly, that my flipping the switch causes the lamp's lighting up also in the original world.

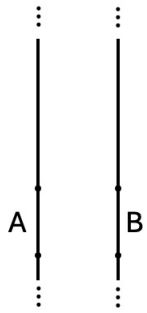
Or consider Threat Cancelation: I turn on a powerful electromagnet, and it deflects a piece of shrapnel that was hurling toward a window. Intuitively, the electromagnet's turning on is a cause of the window's staying intact. The causal past of the window's remaining intact consists, besides the event itself, of the suddenly rising electromagnetic field, and the electromagnet's turning on. But these three events are also present in a world where no piece of shrapnel is hurling toward the window. And in *that* world the electromagnet's turning on is causally unrelated to the window's remaining intact. So Hall's Intrinsicness principle would predict, wrongly, that the electromagnet's turning on likewise isn't a cause of the window's remaining intact in the original world.

These cases undermine Hall's **Intrinsicness** principle, but they don't undermine **Weak Intrinsicness**. The actual temporal history of the lamp's lighting up entails how both

[21] Hall also considers further strengthenings, where $w'$ merely contains a structure "similar" to $S$. These are, of course, subject to the same counterexamples as the current principle.

cables are connected. Likewise, the actual temporal history of the window's remaining intact contains the piece of shrapnel.

It's easy to see how **Weak Intrinsicness** 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 Intrinsicness** that $A = 1$ doesn't cause $B = 1$ in LOOP.

# 3 Troubles for Accounts of Causation

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

So, there can be determinate, non-causal counterfactual dependence, even between positive, proportional, not overly fragile events. What does that mean for counterfactualist reductions of causation? Most obviously, any account which entails **Sufficiency** must be rejected. At least two such accounts come to mind.

The most famous is Lewis (1973a). It says that $c$ causes $e$ iff there is a chain of actual events $d_1, ..., d_n$ with $d_1 = c$ and $d_n = e$ such that, for all $i = 1, ..., n-1$, $(d_i, d_{i+1})$ is a suitable event pair and $\neg O(d_i) \mathrel{\square\!\!\rightarrow} \neg O(d_{i+1})$. In particular, then, if $\neg O(c) \mathrel{\square\!\!\rightarrow} \neg O(e)$ for a suitable pair $(c, e)$ of actual events, $c$ causes $e$, and so Lewis's account entails **Sufficiency**. Our argument against Lewis's account joins the ranks of many previous objections raised against it—notably its failure to handle cases of late preemption and symmetric overdetermination. However, our argument also applies to successor theories which handle these cases.

One of these successors is Hall (2007). According to it, $c$ causes $e$ iff $(c, e)$ is a suitable pair of actual events and there is a "*reduction*" of the actual world in which $c$ counterfactually depends on $e$. It needn't concern us what exactly a reduction is;[22] what matters here is

[22]Just to give a flavor: roughly, it's a situation in which zero or more parts of the world that are actually in a "non-default" (or "deviant") state adopt their default state instead, while the rest is unchanged.

that every world counts as a reduction of *itself* (Hall, 2007, p. 127). Thus we have again, that, where $(c, e)$ is a suitable pair of actual events, $c$'s counterfactually depending on $e$ is sufficient for $c$'s causing $e$—we have, that is, **Sufficiency**. So Hall's (2007) account should be rejected too.[23,24]

## 3.2 Against Structural Equation Accounts

### 3.2.1 The Failure of Counterfactualist Interpretations of Structural Equations

A *structural equations model* (SEM) uses variables to represent sets of events and so-called "structural equations" to represent dependencies between these events. Structural equations are expressions of the form $\ulcorner X := f_X(Y_1, ..., Y_n)\urcorner$, such that $X, Y_1, ..., Y_n$ are variables taking real values and $f_X$ is a function from $n$-tuples of real numbers into real numbers. The equation is interpreted to encode that the values of $Y_1, ..., Y_n$ *determine* $X$'s value—hence the asymmetric symbol ":=".

Formally, a structural equations model 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 on the left-hand side of any structural equation, and every *endogenous* variable appears on the left-hand side of exactly one structural equation. 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.)

To first approximation, SEM accounts of causation say that $X = x$ is an actual cause of $Y = y$ in some given model iff $X = x$ and $Y = y$ are both true and the model contains some relevant circumstances ("contingencies") in which $Y = y$ counterfactually depends on $X = x$. Different SEM accounts of causation differ by what contingencies they consider relevant in evaluating whether $X = x$ causes $Y = y$ in a model. But they agree on a common sufficient condition for causation in a model (cf. Hitchcock (2001), Menzies (2004), Halpern and Pearl (2005), Weslake (2015), and Halpern (2016)). Where $V$ and $W$

[23] In Hall's defense, he is aware of these limitations, 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.

[24] Earlier I discussed Glynn's (2013) account (fn. 4). It's easy to see that it, too, succumbs to our two counterexamples. The counterfactual situation in GAUSS only has "late" miracles anyway—disappearing the electron and changing the field values exactly at $t$—with everything before and after $t$ evolving according to the actual laws. So we straightforwardly have $\neg O(c_G) \blacksquare\!\!\rightarrow \neg O(e_G)$.

As for LOOP, the existence of a global time order is a prerequisite of Glynn's account; so, for argument's sake, let's grant that we can identify times across the two strands of spacetime. Moreover, shrink $A$ to a single point, so that it's part of a single time $t$; and let $B$ occur strictly after $t$. In evaluating $A = 0 \blacksquare\!\!\rightarrow \ldots$ according to Glynn, we then only consider counterfactual worlds with miracles at $t$. But in the closest such world where $A = 0$ we must thus have $B = 0$: by stipulation, there is no miracle after $B$, and hence $A = 0 \wedge B = 1$ would lead to contradiction. So we have $A = 0 \blacksquare\!\!\rightarrow B = 0$, i.e. $\neg O(c_L) \blacksquare\!\!\rightarrow \neg O(e_L)$. So Glynn's account wrongly entails action at a distance in both GAUSS and LOOP.

are variables, a *causal path from $V$ to $W$* is a sequence of variables $(X_1, ..., X_n)$ such that $V = X_1$ and $W = X_n$ and, for all $i = 1, ..., n-1$, the value of $f_{i+1}$ depends on $X_i$—that is, $X_i$ appears ineliminably on the right-hand side of $X_{i+1}$'s structural equation. What the different SEM accounts agree on is that holding all "off-path" variables fixed at their *actual* values is a relevant contingency.

Now, as others have pointed out (cf. Hall (2007)), reductions of causation can't stop merely at model-relative definitions of causation. My letting go of the pen doesn't just cause it to fall relative to one or another model—it causes it to fall *full stop*. Any serious account of causation should reproduce this judgment. SEM accounts of causation thus introduce 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 common denominator of SEM accounts of causation can now be put thus:

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

A satisfactory account of causation must clarify the notion of adequacy—doubly so if the account is a reduction of causation. A minimal requirement on adequacy is that adequate SEMs contain only *true* (or at least approximately true) structural equations (Hitchcock, 2001, p. 292). Hitchcock (2001), Menzies (2004), Halpern and Pearl (2005), and Weslake (2015) all agree that the truth conditions of structural equations are given in terms of *counterfactual conditionals*. 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 $Z = f_Z(x, y, ..., w)$."

Similarly, Menzies (2004, p. 822):[25]

> "[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."

[25] Curiously, Menzies uses indicative conditionals here, even though he means them to be "counterfactuals". I've thus substituted the subjunctive form.

Halpern and Pearl (2005, p.847):

> "[Structural equations] support a counterfactual interpretation. For example, the equation $[X := Y + U]$ claims that in the context $U = u$, if $Y$ were 4, then $X$ would be $u + 4$".

Finally, Weslake (2015):

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

Besides containing only true (or approximately true) structural equations, an adequate causal model should also be (what one might call) *minimal* but *exhaustive*.[26] Intuitively, a SEM should contain enough causal paths to capture all counterfactual dependencies, but no more than necessary. Of the above authors, Hitchcock (2001) develops these two conditions most explicitly.

- **Minimality:** An adequate SEM $\mathcal{M}$ is *minimal*: i.e., if, for every combination of values of $x, x', z, y, ..., w$ of the endogenous variables $X, Y, Z, ..., W$ in $\mathcal{M}$,

$$\begin{aligned}(X = x \wedge Y = y \wedge ... \wedge W = w \mathrel{\square\!\!\rightarrow} Z = z) \\ \leftrightarrow (X = x' \wedge Y = y \wedge ... \wedge W = w \mathrel{\square\!\!\rightarrow} Z = z),\end{aligned} \tag{3}$$

  then the value of the right-hand side of $Z$'s structural equation in $\mathcal{M}$ is independent of $X$'s value.[27]

- **Exhaustiveness:** An adequate SEM $\mathcal{M}$ is *exhaustive*: i.e., if, for some combination of values $x, x', y, z, ..., w$ of the endogenous variables $X, Y, Z, ..., W$ in $\mathcal{M}$,

$$\begin{aligned}(X = x \wedge Y = y \wedge ... \wedge W = w \mathrel{\square\!\!\rightarrow} Z = z) \\ \wedge \neg (X = x' \wedge Y = y \wedge ... \wedge W = w \mathrel{\square\!\!\rightarrow} Z = z),\end{aligned} \tag{4}$$

[26] Additionally there are constraints on variable choice: where $X, Y \in \mathbf{U} \cup \mathbf{V}$ are distinct variables, the pair $(X = x, Y = y)$ of events should be *suitable*. That is, $X = x$ and $Y = y$ should be "distinct", in Lewis's (1973; 1979) sense; additionally, one might require that they be proportional to each other (cf. Halpern and Hitchcock, 2010, Sec. 3); that they be "positive" events; and that they not be overly fragile. The following models of GAUSS and LOOP satisfy all of these constraints.

[27] Hitchcock's original quote, in full (where $\mathcal{E}_{//}$ are equations with an endogenous variables on their left-hand side): "Equations in $\mathcal{E}_{//}$ must always be written in minimal form: if for all $x, x', y, z, ..., w$, $f_Z(x, y, ..., w) = f_Z(x', y, ..., w)$, then the value of $Z$ does not depend upon the value of $X$ at all, and the structural equation for $Z$ must be rewritten $Z = f_Z(Y, ..., W)$" (Hitchcock, 2001, p. 280). It is clear from context that by "$f_Z(x, y, ..., w) = f_Z(x', y, ..., w)$" Hitchcock means our condition 3.

then the value of the right-hand side of $Z$'s structural equation in $\mathcal{M}$ depends on $X$'s value.[28]

Hitchcock's conditions involve evaluating counterfactuals which explicitly suspend the underlying causal structure. For example, suppose a model consists of the variables $\{X, Y, Z\}$, and includes the structural equation $Z := \max(X, Y)$. Now suppose that, in fact, $X = 1$ is the sole cause of $Y = 1$. Nonetheless, to evaluate whether the model, including $Z := \max(X, Y)$, is adequate we'll have to evaluate the counterfactual "If $X = 1$, but $Y \neq 1$ anyway, then..."; the antecedent explicitly suspends the causal connection from $X$ to $Y$. Hitchcock (2001, p. 275) calls these counterfactuals "*explicitly nonforetracking*" (ENF) counterfactuals.

So, the two common elements in SEM accounts are **Sufficiency**$^*$ and a counterfactualist notion of adequacy, with the conjunction of **Minimality** and **Exhaustiveness** being the most developed expression of the latter. Unfortunately, this combination fails for both GAUSS and LOOP.

For GAUSS, consider a model $\mathcal{M}_G = (\mathbf{U}_G, \mathbf{V}_G, \mathcal{E}_G)$ with variables $\mathbf{U}_G \cup \mathbf{V}_G = \{E, F\}$. By **Dependence$_\mathbf{G}$**, the value of $F$ counterfactually depends on the value of $E$. But $E$ and $F$ are the only variables in the model, and so, by **Exhaustiveness**, $F$'s structural equation must contain $E$ if $\mathcal{M}_G$ is to be adequate. Indeed, since $E = 1 \mathrel{\Box\!\!\rightarrow} F = 1$ and $E = 0 \mathrel{\Box\!\!\rightarrow} F = 0$, the structural equation must be

$$F := E. \tag{5}$$

But now **Sufficiency**$^*$ immediately entails that $E = 1$ causes $F = 1$, contradicting **Non-Causation$_\mathrm{G}$**. (It may or may not also be true that $F = 1 \mathrel{\Box\!\!\rightarrow} E = 1$ and $F = 0 \mathrel{\Box\!\!\rightarrow} E = 0$, and hence that the converse of eq. 5, $E := F$ should be part of $\mathcal{M}_G$. In any case, it doesn't matter for my argument.)

Similarly for LOOP: consider a model $\mathcal{M}_L = (\mathbf{U}_F, \mathbf{V}_F, \mathcal{E}_F)$ with variables $\mathbf{U}_F \cup \mathbf{V}_F =$

[28] Hitchcock's original quote, in full: "By the same token, equations in $\mathcal{E}_{//}$ must always include as arguments any variables in $\mathcal{V}$ upon which $Z$ counterfactually depends, given the values of the other variables. If, for some $x, x', y, z, ..., w$, $f_Z(x, y, ..., w) \neq f_Z(x', y, ..., w)$, then the value of $Z$ does depend upon the value of $X$, and $Z = f_Z(Y, ..., W)$ is not in $\mathcal{E}_{//}$. The correct equation for $Z$ can be arrived at by expressing the value of $Z$ as a function of *all* other variables in $\mathcal{V}$ and then eliminating those variables whose values are redundant given every assignment of values to the other variables." (p. 281)

The condition "$f_Z(x, y, ..., w) \neq f_Z(x', y, ..., w)$" can arguably be read either as condition 4 or as the condition

$$(X = x \& Y = y \& ... \& W = w \mathrel{\Box\!\!\rightarrow} Z = z) \wedge (X = x' \& Y = y \& ... \& W = w \mathrel{\Box\!\!\rightarrow} Z \neq z).$$

The conditions are equivalent given Conditional Excluded Middle; otherwise condition 4 is weaker. Nothing substantial hinges on which condition we choose.

$\{A, B\}$. Since the value of $A$ counterfactually depends on the value of $B$ (as per **Dependence$_{\mathbf{L}}$**) and $A$ and $B$ are the only variables in the model, $B$'s structural equation must, by **Exhaustiveness**, contain $A$ if $\mathcal{M}_L$ is to be adequate. Indeed, since $A = 1 \,\square\!\!\rightarrow B = 1$ and $A = 0 \,\square\!\!\rightarrow B = 0$, the structural equation must be

$$B := A. \tag{6}$$

But now **Sufficiency**$^*$ immediately entails that $A = 1$ causes $B = 1$, contradicting **Non-Causation$_{\mathrm{L}}$**. (We also have $B = 1 \,\square\!\!\rightarrow A = 1$ and $B = 0 \,\square\!\!\rightarrow A = 0$, and hence that the converse of eq. 6, $B := A$ should be part of $\mathcal{M}_L$.)

### 3.2.2 Impoverished Models?

One might worry that our two arguments rely on impoverished models, with too few variable values, and that those models can be dismissed on independent grounds as inadequate. However, the concern can't be that models with few variable values are categorically inadequate. Many canonical examples of adequate causal models use only a small number of variables and values—e.g., cases of causal overdetermination may be modeled with just three binary variables, and cases of linear causation between two variables with only two.

Instead, the concern is that there is some case-dependent minimum threshold for detail of a causal model. The need for such a threshold is known from late preemption and overdetermination cases (cf. Halpern and Pearl, 2005; Handfield et al., 2008; Halpern and Hitchcock, 2010). Consider a classic example of late preemption: Susy and Billy throw rocks at a bottle. Billy's rock arrives first, smashing the bottle; thus, Billy's throw is a cause of the bottle's shattering, while Susy's throw is not. One might naively try to model this scenario with *three* binary variables, $S$, $B$, $D$, representing, respectively, Susy's throw, Billy's throw, and the bottle's shattering. Further, one might posit a single structural equation,

$$D := \max(S, B),$$

capturing the fact that each throw on its own is already sufficient for the bottle's shattering. But the resulting SEM is identical to one of symmetrical overdetermination, i.e. a scenario in which Susy's and Billy's rocks arrive at *exactly the same time*. In *that* scenario, either both throws cause the bottle's shattering, or neither does (opinions vary on this). So SEM accounts of causation yields the wrong results for at least one of the two scenarios.

Halpern and Pearl (2005) provide the following diagnosis of the problem:

> "If we want to argue in a case of preemption that $X = x$ is the cause of $\varphi$ rather than $Y = y$, then there must be [an additional] random variable ... that takes on different values depending on whether $X = x$ or $Y = y$ is the actual cause." (p.862)

This diagnosis seems correct to me, but it cannot itself be part of a reductive definition of adequacy, because it primitively appeals to causation, twice over: once in diagnosing a problem as one of *preemption*, and once in assessing the suitability of the additional variable. (Handfield et al. (2008, p. 153-4) make a similar point.)

As an additional reply, Halpern and Pearl (2005, pp. 863-4) note that fine-graining events by occurrence time can transform cases of preemption into ordinary cases of inhibition—cases which their account handles better. Whatever its merits as a solution to the problem of preemption, the suggestion helps neither with GAUSS nor with LOOP, where the problems persist even if we focus on events occurring at single times (or small intervals of time).

Halpern and Hitchcock (2010) propose that, in adequate models, adding further variables shouldn't change the 'topology' of the model. But no precise definition of the notion is provided, and it remains unclear if there exists one free of primitive causal language. Moreover, the only example of a 'topology' change they *do* provide are the addition of "cross" causal connections (for example, in the above preemption case, a connection from an additional variable $S'$ downstream from $S$ to an additional variable $B'$ downstream from $B$). But this would make adequacy way too demanding: for virtually *any* causal model, there are some additional variables we could add that would engender cross causal connections. For consider variables representing *microscopic* intermediate events: such events are easily influenced by other "off-branch" events, therefore necessitating "cross" causal connections.

But suppose, for argument's sake, that the SEM advocate *did* find principled grounds to exclude $\mathcal{M}_G$ and $\mathcal{M}_L$ as inadequate. It wouldn't help: even if we restrict ourselves to enriched models, the combination of **Sufficiency**$^*$ and **Exhaustiveness** yields false predictions about GAUSS and LOOP.

Consider enriching GAUSS by adding a sufficient common cause of $E = 1$ and $F = 1$: the state of the world at a time $t^-$ where the electron is located inside the past light-cone of $S$-at-$t$. See fig. 3 for an illustration. Let the binary variable $C$ represent this common cause, with $C = 1$ indicating that the state at $t^-$ is as indicated in the figure: in particular, that the electron is present at $x$ at $t^-$ and that the net electric flux through the surface $S^-$ is negative. Let $C = 0$ indicate that the electron is absent at $x$ at $t^-$ and that the net electric flux through $S^-$ is zero. Now consider a richer model $\mathcal{M}^*_G = (\mathbf{U}^*, \mathbf{V}^*, \mathcal{E}^*)$ with $\mathbf{U}^* \cup \mathbf{V}^* = \{C, E, F\}$.

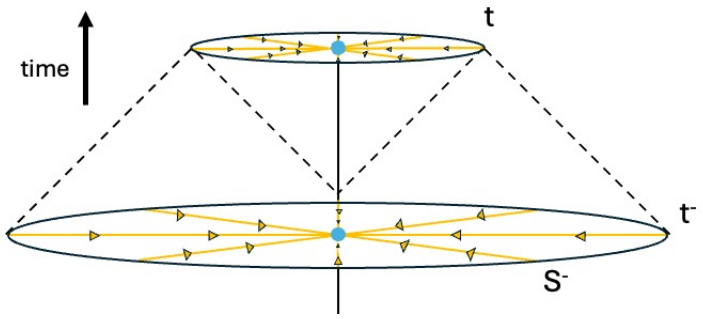



Figure 3: A (2+1)D sketch (two spatial dimensions plus one temporal dimension)

*Actually*, $C = i$ is a cause of $E = i$ and of $F = i$ (for both $i = 0, 1$). Moreover, to avoid the previous trouble, $F$'s and $E$'s structural equations should be independent of $E$ and $F$, respectively. Hence $\mathcal{M}^*_G$ should contain the following two structural equations:

$$E := C, \tag{7}$$

$$F := C. \tag{8}$$

Moreover, since neither of $E$ or $F$ causes $C$, these should be the *only* structural equations (i.e., $C$ should be an exogenous variable in $\mathcal{M}^*_G$). Does Hitchcock's definition of adequacy vindicate the resulting model as adequate?

No. The right-hand side of eq. 8 is independent of $E$. Moreover, by And-to-If[29] we have $C \wedge E \mathrel{\Box\!\!\rightarrow} F$. Hence, by **Exhaustiveness**, the inclusion of eq. 8 requires the truth of the following conditional:

$$C \wedge \neg E \mathrel{\Box\!\!\rightarrow} F, \tag{9}$$

But, by **Dependency**$_G$, we have

$$\neg E \mathrel{\Box\!\!\rightarrow} \neg F. \tag{10}$$

Now, the following rule is valid in any standard counterfactual logic:[30]

[29] I.e., the rule $X \wedge Y \vdash X \mathrel{\Box\!\!\rightarrow} Y$, valid in any standard logic of counterfactuals, including Stalnaker (1968) and Lewis (1973b).

[30] The rule (sometimes called "Possibility Transfer") is entailed by propositional logic, Rational Monotonicity (RM)—i.e. $A \mathrel{\Diamond\!\!\rightarrow} B, A \mathrel{\Box\!\!\rightarrow} C \vdash (A \wedge B) \mathrel{\Box\!\!\rightarrow} C$—plus the two rules (*) $\Box(A \supset B) \vdash A \mathrel{\Box\!\!\rightarrow} B$, and (**) $\Diamond A, A \mathrel{\Box\!\!\rightarrow} C \vdash A \mathrel{\Diamond\!\!\rightarrow} C$. (Here, $\cdot \mathrel{\Diamond\!\!\rightarrow} \cdot$ is the dual of $\mathrel{\Box\!\!\rightarrow}$, i.e. $\cdot \mathrel{\Diamond\!\!\rightarrow} \cdot \equiv \neg(\cdot \mathrel{\Box\!\!\rightarrow} \neg\cdot)$.) All of these rules are part of any standard logic of counterfactual conditionals, including Stalnaker (1968) and Lewis (1973b).
*Proof of the entailment:* An instance of one of RM's contrapositives is this:

$$X \mathrel{\Box\!\!\rightarrow} \neg Z, X \wedge Y \mathrel{\Box\!\!\rightarrow} Z \vdash X \mathrel{\Box\!\!\rightarrow} \neg Y. \tag{13}$$

But eqs. 9, 10, and 13 together entail[31]

$$\neg E \mathrel{\Box\!\!\rightarrow} \neg C. \tag{14}$$

On the standard reading, this counterfactual is false. Worse, its truth would commit SEM accounts to the absurd conclusion that the electron's presence at $x$ at $t$ *causes* the electron's presence at $t^-$, given the adequacy of models with variables $E$ and $C$ (plus optionally any number of other variables in $C$'s past).

(Moreover, given Conditional Excluded Middle (CEM)—i.e., the rule $\vdash X \mathrel{\Box\!\!\rightarrow} Y \vee X \mathrel{\Box\!\!\rightarrow} \neg Y$, valid in Stalnaker's logic and defended by several others, e.g. by Stalnaker (1981), Williams (2010), and Goodman (ms.)—we can formally *prove* the inadequacy of $\mathcal{M}^*_G$. For by 10 and 14 we have $\neg E \wedge \neg F \mathrel{\Box\!\!\rightarrow} \neg C$. If $E \wedge \neg F \mathrel{\Box\!\!\rightarrow} C$, **Exhaustiveness** thus requires that $C$'s structural equation depend on $E$, and so $\mathcal{M}^*_G$ is inadequate. If instead $\neg(E \wedge \neg F \mathrel{\Box\!\!\rightarrow} C)$, then by CEM, $E \wedge \neg F \mathrel{\Box\!\!\rightarrow} \neg C$. But, by And-to-If, we have $E \wedge F \mathrel{\Box\!\!\rightarrow} C$. But now **Exhaustiveness** requires that $C$'s structural equation depend on $F$, and so $\mathcal{M}^*_G$ is, again, inadequate.)

So, the given counterfactualist reduction of adequacy cannot realistically vindicate $\mathcal{M}^*_G$, which lacks a causal paths from $E$ to $F$, as adequate. It's hard to see how this could be different for other enriched models for GAUSS which contain $E$ and $F$.

The situation is similarly straightforward with LOOP. Enriching $\mathcal{M}_L$ by introducing additional variables to the past of $A$ and $B$ won't eliminate the equation $B := A$ from the model. Alternatively, one may also introduce additional variables, $A^+$ and $B^+$, to the future of $A$ and $B$, respectively, closer to the three-way fork. In the enriched model, $B$'s structural equation no longer depends on $A$, for holding fixed $A^+$ and $B^+$'s values, $A$ and $B$ are counterfactually independent. But now $A^+$ and $B^+$ are themselves counterfactually

---

$$X \mathrel{\Box\!\!\rightarrow} \neg Z, X \wedge Y \mathrel{\Diamond\!\!\rightarrow} Z \vdash X \mathrel{\Box\!\!\rightarrow} \neg Y. \tag{11}$$

But $\vdash \Box\neg(X \wedge Y) \vee \Diamond(X \wedge Y)$ in propositional logic. By (*), $\Box\neg(X \wedge Y) \vdash X \mathrel{\Box\!\!\rightarrow} \neg Y$. By (**), $\Diamond(X \wedge Y) \vdash (X \wedge Y \mathrel{\Box\!\!\rightarrow} Z \supset X \wedge Y \mathrel{\Diamond\!\!\rightarrow} Z)$. Hence

$$\vdash (X \mathrel{\Box\!\!\rightarrow} \neg Y) \vee (((X \wedge Y) \mathrel{\Box\!\!\rightarrow} Z) \supset ((X \wedge Y) \mathrel{\Diamond\!\!\rightarrow} Z)). \tag{12}$$

But 11 and 12 together imply

$$X \mathrel{\Box\!\!\rightarrow} \neg Z, X \wedge Y \mathrel{\Box\!\!\rightarrow} Z \vdash X \mathrel{\Box\!\!\rightarrow} \neg Y. \blacksquare$$

[31] To see this instantiate, in 13, $X$ with $\neg E$, $Y$ with $C$, and $Z$ with $F$.

interdependent, even holding fixed the values of the remaining variables $X, Y, A, B$. Now, you may introduce yet additional variables, downstream from $A^+$ and $B^+$—but the same game repeats. Whatever final two intervals of the $A$-branch and the $B$-branch the model represents, the corresponding variables will be counterfactually interdependent.[32] So given **Exhaustiveness**, any adequate enrichment of $\mathcal{M}_L$ must include a causal path from somewhere on the $A$-branch to somewhere on the $B$-branch and therefore, given **Sufficiency**$^*$, yield that the $A$-branch's field values cause the $B$-branch's field values—which is false.[33]

# 4 The Path Forward

In the presence of non-dynamical nomic constraints, counterfactual dependence ceases to track causal dependence. As a result, I've argued, extant counterfactualist reductions of causation are false. By contrast, *dynamical* nomic constraints—*viz.*, dynamical laws, like Newton's law, the classical Klein-Gordon equation, or Schrödinger's equation——do seem to support causation. They describe how physical states evolve through time: given the state of the world at one time, dynamical laws generate future trajectories of the universe. (In the deterministic case, they also determine a past trajectory, and in the stochastic case, they instead generate a probability distribution over possible future trajectories.) This

[32] Some models with infinitely many variables might lack a *final* represented interval on the $A$-branch and/or $B$-branch. But it's extremely implausible that *only* such models are adequate. Indeed, since any such model lacks any causal paths either from $A$ to $X$ or from $B$ to $X$, if they *were* the only adequate models, SEM accounts would judge that neither $A = 1$ nor $B = 1$ causes $X = 1$, which is absurd.

[33] Gallow's (2016) more sophisticated counterfactualist theory of adequacy also fails in cases of non-dynamical nomic constraints. (He brackets the cyclic case in his discussion—but, again, we can't afford this if our goal is an analysis of causation.) Let $\phi$ be the selection function for your favorite semantics of counterfactual conditionals, mapping proposition-world pairs into sets of worlds. For any world $w$ and any given set of variables $\mathbf{X}$, let the $\mathbf{X}$-*closure of* $w$ *under* $\phi$ be the closure of $\{w\}$ under the set of functions $\{\phi(\mathbf{X}' = \mathbf{x}', \cdot) | \mathbf{X}' \subseteq \mathbf{X}$ and $\mathbf{x}$ is in the range of $\mathbf{X}\}$, i.e., the smallest set $W$ such that: (i) $w \in W$ and (ii) if $w' \in W$ and $\mathbf{x}'$ is in the range of some subset $\mathbf{X}' \subseteq \mathbf{X}$, then $\phi(\mathbf{X}' = \mathbf{x}', w') \subseteq W$. Intuitively, the $\mathbf{X}$-*closure of* $w$ *under* $\phi$ is exactly the set of worlds you can reach from $w$ by repeatedly subjunctively supposing $\mathbf{X}' = \mathbf{x}'$—i.e., taking conditionals of the form $\mathbf{X}' = \mathbf{x}' \mathrel{\square\!\!\rightarrow} ...$—where $\mathbf{X}'$ is a subset of $\mathbf{X}$. Now, according to Gallow, given a selection function $\phi$, an adequate SEM $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathcal{E})$ contains a structural equation $(V := f_V(\mathbf{W})) \in \mathcal{E}$ only if the (ordinary) equation $V = f_V(\mathbf{W})$ is true throughout the actual world's $\mathbf{U} \cup \mathbf{V} \backslash \{V\}$-closure under $\phi$. The "only if" is strengthened to a biconditional if additionally all variables in $\mathbf{U}$ are mutually counterfactually independent throughout that closure (formally, if no SEM with the same variable set but "strictly more" determination relations (i.e., directed causal paths) satisfies the aforementioned property).

According to this semantics, $\mathcal{M}_G$, if it is to be an adequate SEM for GAUSS, must contain the structural equation 5, i.e. $F := E$. However many subjunctive suppositions of the form $E = i \mathrel{\square\!\!\rightarrow} \ldots$, for $i = 0, 1$, are nested, Gauss's law would still hold at all times after $t$. Similarly for $\mathcal{M}_L$: however many subjunctive suppositions of the form $A = i \mathrel{\square\!\!\rightarrow} \ldots$, for $i = 0, 1$, are nested, the dynamics and topological structure downstream from $A$ and $B$ would be unchanged and so we'd still have that $A = B$. Analogous things hold for enriched models of GAUSS and LOOP. Gallow's theory of adequacy does no better than Hitchcock's when faced with non-dynamical nomic constraints.

generation, I suggest, supports causation.

*Local* dynamical laws—as e.g. the classical Klein-Gordon equation, or Newton's equations with a local force law—generate along spacetime structure: the state of a spacetime region is generated from the state of any final section of its past light-cone. If dynamical laws ground causation, one may therefore require that, in adequate models, the structural equation of a variable $W$ depend on $V$ only if there are values $v$ and $w$ such that $V = v$ is in the past light-cone of $W = w$, and there is at least one causal curve from $V = v$ to $W = w$ which is not intersected by another event in the model.

On this vision, SEM adequacy is reducible to dynamical laws and spacetime structure. We can see the vision's promise when looking at GAUSS and LOOP: because $F = 0, 1$ are outside of the light-cones of $E = 0, 1$, the criterion immediately renders inadequate any model with a structural equations for $F$ that depend on $E$. Similarly, since $B = 0, 1$ are outside of the light-cones of $A = 0, 1$, no model with a structural equations for $B$ that depend on $A$ will be adequate. So we avoid the problematic equations which plagued **Exhaustiveness**.

I will undertake this reduction in the next chapter. For now suppose it succeeds, and that we recover the intuitively correct models for GAUSS and LOOP, with no causal paths between dynamically independent variables. We can combine this new reductive account of SEMs with extant reductions of causation to SEMs. Does this yield a satisfactory final reduction of causation? As I show elsewhere in my dissertation, the answer is no. Halpern and Pearl (2005)—arguably the most famous and influential SEM account of causation—fails because their account doesn't make the existence of a causal path between variables a necessary condition for causation. (Their proposal is significantly stronger than **Sufficiency**$^*$.) Other accounts—including Hitchcock (2001), Menzies (2004), and Weslake (2015)—do, and as a result fare a bit better.[34] But they, too, ultimately fail: they still generally produce wrong results for cyclic models, where value assignments to exogenous variables can have zero or multiple solutions. Moreover, I show that no simple modifications of their accounts (such as swapping a universal quantifier for an existential quantifier, or including a non-vacuity condition) solves the problem. Work remains.

[34]The importance of "path-sensitivity"—of making the existence of a causal path between variables necessary for causation—has hitherto gone largely 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 path-*sensitive* accounts. (Though Weslake adds a clarification in a footnote.)

# 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.

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 (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.

Goodman, Jeremy (ms.). "Consequences of Conditional Excluded Middle".

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 (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.

Handfield, Toby et al. (2008). "The Metaphysics of Causal Models: Where's the Biff?" In: *Erkenntnis (1975-)* 68.2, pp. 149–168.

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.

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.

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

— (1986b). "Postscripts to 'Causation'". In: *Philosophical Papers Vol. Ii*. Ed. by David Lewis. Oxford University Press.

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 C. (1981). "A Defense of Conditional Excluded Middle". In: *IFS: Conditionals, Belief, Decision, Chance and Time*. Ed. by William L. Harper, Robert Stalnaker, and Glenn Pearce. Dordrecht: Springer Netherlands, pp. 87–104.

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".

— (2017). "Difference-Making, Closure and Exclusion". In: *Making a Difference: Essays on the Philosophy of Causation*. Ed. by Helen Beebee, Christopher Hitchcock, and Huw Price. Oxford University Press, pp. 215–231. (Visited on 10/31/2024).

Williams, J. Robert G. (2010). "Defending Conditional Excluded Middle". In: *Noûs* 44.4, pp. 650–668.

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.