The Epochetic Analysis of Causation Compared to Counterfactual Accounts

Holger Andreas[1] and Mario Günther*[2]

[1]University of British Columbia
[2]LMU Munich & Carnegie Mellon University

This chapter compares Andreas and Günther's (forthcoming) epochetic analysis of actual causation to the currently popular counterfactual accounts. The primary focus will be on the shortcomings of the counterfactual approach to causation. But we will also explain the motivation behind counterfactual accounts and how the counterfactual approach has successively moved away from its core idea in response to recalcitrant counterexamples. The upshot is that our epochetic analysis tallies better with our causal judgments than the counterfactual accounts.

A comparison to counterfactual accounts at manageable length must be selective. For reasons of systematicity, we have chosen Lewis's (1973a) analysis of causation in terms of chains of difference-making, Yablo's (2002) account in terms of de facto dependence, and the causal model accounts of Hitchcock (2001), Halpern and Pearl (2005), Halpern (2015), Halpern (2016), and Gallow (2021). The latter may be seen as the current culmination of the counterfactual approach and the strongest competitor to our epochetic analysis. This is why we devoted a rather long section on Gallow's theory towards the end of this chapter.

*Mario.Guenther@lmu.du

Contents

1 Counterfactual Dependence

The motivating idea and starting point of many counterfactual accounts is that counterfactual dependence between distinct occurring events is sufficient for causation.[1] An event E counterfactually depends on an event C just in case the counterfactual conditional "if C had not occurred, then E would not have occurred" is true. The event C is thus a cause of the distinct event E when C and E occur, and E would not have occurred, had C not occurred. Suppose Suzy throws a rock at a window and the window shatters. If the counterfactual conditional "the window would not have shattered if Suzy had not thrown the rock" is true, then Suzy's throwing the rock is a cause of the window's shattering.

Counterfactual dependence between distinct occurring events should, however, not be elevated to a necessary and sufficient condition for causation. This elevation would result in the *simple counterfactual account*:

> An event C is a cause of a distinct event E iff
>
> (1) C and E occur, and
>
> (2) if C had not occurred, E would not have occurred.

There are direct counterexamples to the simple counterfactual account, namely when an effect would occur even if one of its genuine causes would have been absent. In such cases of redundant causation, there is more than one event that would be sufficient for the effect to occur.

One type of troublesome redundant causation is overdetermination. Let's say Suzy and Billy each throw a rock at a window, the rocks impact upon the window at the same time and each rock alone would have been sufficient to break the window. Each rock throwing is arguably a cause of the window's breaking. But the simple counterfactual account says that neither Suzy's nor Billy's throwing is a cause of the window's shattering. Had Suzy not thrown her rock, the window would have shattered

[1] See Lewis (1973a, 2000); Ramachandran (1997); Hitchcock (2001); Yablo (2002, 2004); Woodward (2003); Hall (2004, 2007); Halpern and Pearl (2005); Halpern (2015), and many others.

anyways—due to Billy's throw. And had Billy not thrown his rock, the window would have shattered anyways—due to Suzy's throw. But then, what caused the shattering of the window? Surely, we do not want to say that the shattering is uncaused. We will come back to this point.

Another type of troublesome redundant causation is early preemption. Let's say, again, that Suzy and Billy each throw a rock at a window. But this time, Suzy's rock deflects Billy's mid-flight. Only Suzy's rock impacts upon the window and it shatters. Had Suzy not thrown, however, Billy's rock would not have been deflected and would have shattered the window. For convenience, we reproduce here the neuron diagram exhibiting the canonical structure of early preemption:

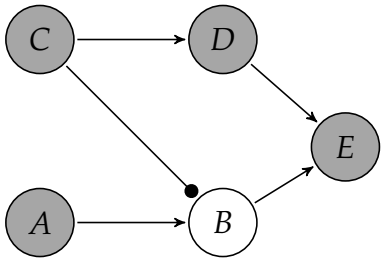

Figure 1: Early preemption

Suzy throws her rock (C) towards the window, and Billy his (A). Suzy's rock impacts upon the window (D), and prevents Billy's rock from doing so ($\neg B$) by deflecting it. The impact of Suzy's rock (D) shatters the window (E). Had Suzy not thrown her rock ($\neg C$), however, Billy's rock (A) would have impacted upon the window (B), which would have shattered the window (E). The simple counterfactual account thus says that Suzy's throw is not a cause of the window's shattering. But Suzy's throw is the actual cause of the window's shattering, one that preempts Billy's throw—a mere would-be cause of the shattering. We take the simple counterfactual account to be refuted.[2]

[2]Coady (2004) defends the simple counterfactual account. We explain in Andreas and Günther (2025) why we think his defence fails.

We have seen that redundant causation means trouble for the simple counterfactual account. And no wonder. An effect does not counterfactually depend on a redundant cause. Less simple counterfactual accounts thus drop the necessity of counterfactual dependence for causation and so deviate from the idea that causation is nothing but counterfactual dependence between actual events. In the next sections, we will turn to such less simple counterfactual accounts. We begin with Lewis's (1973a) analysis of causation whose importance for current counterfactual accounts is hard to overstate.[3]

2 Difference-Making Chains

Lewis (1973a, p. 557) proclaims that we think of causes as difference makers. Whether or not a cause occurs makes a difference as to whether or not its effect occurs. He spells out this idea of difference making in terms of subjunctive conditionals. If Suzy had thrown her rock, the window would have shattered. And if Suzy had not thrown her rock, the window would not have shattered. Suzy's throw makes a difference to the window's shattering just in case both subjunctives are true. In general:

> An event or absence C makes a difference to a distinct event or absence E iff two subjunctive conditionals are true:
>
> (i) if C had occurred, E would have, and
>
> (ii) if C had not occurred, E would not have.

Lewis's (1973b) semantics says that the subjunctive (i) is true whenever C and E occur. Hence, C makes a difference to E if C causes E on the simple counterfactual account.

Perhaps surprisingly, Lewis does not say that an event is a cause *in virtue of* making a difference to another. He rather analyses causation as the transitive closure of difference-making which we may sum up as follows:

[3] Section 2 and 9 of this chapter draw on material from Andreas and Günther (2025).

> An event or absence C is a cause of a distinct event or absence E iff there is a difference-making chain running from C to E.

A difference-making chain is a finite sequence of distinct actual events and absences such that each element in the sequence makes a difference to its successor. In symbols, a finite sequence $\langle C, D_1, ..., D_n, E\rangle$ of distinct actual events and absences is a difference-making chain from C to E iff C makes a difference to D_1, D_1 makes a difference to ..., and D_n makes a difference to E. So Lewis does—strictly speaking—not think of causes as difference-makers. He thinks of causes as *initiators of difference-making chains*.

Difference-making suffices for causation. Suppose the event C makes a difference to the distinct event E. Then there is a finite sequence $\langle C, E\rangle$ of distinct actual events such that each element in the sequence makes a counterfactual difference to its successor. Hence, C is a cause of E on Lewis's (1973a) analysis.

By contrast, causation does not suffice for difference-making on Lewis's analysis. Suppose an event C makes a counterfactual difference to another event D, which in turn makes such a difference to a third event E. Then C is a cause of E—even if C does not make a difference to E. A case in point is the early preemption scenario depicted in Figure 1. Suzy's throw of a rock makes a difference as to whether or not her rock impacts upon the window. And given that Suzy's rock deflected Billy's, her rock impacting upon the window makes a difference as to whether or not the window shatters. Still, Suzy's throw does not make a difference as to the window's shattering. Had she not thrown her rock, Billy's rock would not have been deflected and so would have shattered the window. Difference-making is not transitive. But Lewis thinks causation is, and so defines it to be transitive.

Lewis's analysis solves the early preemption scenario. As we have just seen, Suzy's throw (C) counts as a cause of the window's shattering (E). C and E occur, and there is the sequence $\langle C, D, E\rangle$ of distinct events such that the counterfactuals $\neg C \mathrel{\Box\!\!\rightarrow} \neg D$ and $\neg D \mathrel{\Box\!\!\rightarrow} \neg E$ are true. By contrast, there is no sequence of distinct actual events and absences from Billy's throw (A) to the window's shattering (E) such that each element makes a difference to its successor. If A had not occurred, B still would not have

occurred. A does not initiate a counterfactual difference-making chain to E.

Note that Lewis's solution to the problem of early preemption relies on the assumption that backtracking counterfactuals are false. Without this assumption, $\neg D \mathrel{\Box\!\!\rightarrow} \neg E$ would not come out true. For this to be seen, suppose counterfactual conditionals going against the direction of time and causation may come out true. Then the conditional "if Suzy's rock had not impacted upon the window, she would not have thrown her rock" is true ($\neg D \mathrel{\Box\!\!\rightarrow} \neg C$). And so the window would have shattered (E) because Billy's rock would have touched the window (B) had it not been deflected by Suzy's rock ($\neg C$). Taken together, the window would have shattered if Suzy's rock had not touched the window ($\neg D \mathrel{\Box\!\!\rightarrow} E$). Hence, there is no chain of difference-making from C to E when backtracking counterfactuals may be true. For a closer look at backtracking counterfactuals and Lewis's semantics of counterfactuals, we refer the reader to Andreas and Günther (forthcoming, Ch. 10, Sec. 7).

Lewis's solution to early preemption fails to work for scenarios of late preemption, schematically depicted by Figure 2:

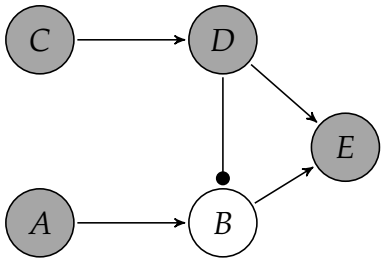

Figure 2: Late preemption

Suppose, for contradiction, that Lewis's analysis works for scenarios of late preemption. This implies that C comes out as a genuine cause of E. From this we can infer that the counterfactuals $\neg C \mathrel{\Box\!\!\rightarrow} \neg D$ and $\neg D \mathrel{\Box\!\!\rightarrow} \neg E$ are true. Applied to the rock-throwing example, $\neg D \mathrel{\Box\!\!\rightarrow} \neg E$ means that the window would not have shattered, had Suzy's rock not

touched it. But this counterfactual is false, even if backtracking is excluded. Had Suzy's rock not hit the bottle, Billy's would have, and so the window would have shattered anyways.

Scenarios of overdetermination are troublesome too. As is well-known, Lewis's analysis does not count symmetric overdeterminers as causes. Unlike preemption, there is no causally relevant difference between overdeterminers. Lewis's analysis delivers the same verdict as the simple counterfactual account: neither Suzy's throw nor Billy's throw counts as a cause of the window's shattering in the overdetermination scenario—a verdict which strikes many as wrong.

Lewis (1986, pp. 199n), however, thinks it is only clear that the symmetric overdeterminers are on a par:

> It may or may not be clear whether either [overdeterminer] is a cause; but it is clear at least that their claims are equal. There is nothing to choose between them. Both or neither must count as causes.

To decide whether both throws or neither should count as a cause is for Lewis up to our best theory of causation. It is "spoils to the victor for lack of firm common-sense judgements." (p. 208) Woodward (2003, p. 85) counters: "My guess is that Lewis is wrong about common sense." Or in the words of Paul and Hall (2013, p. 152): "It seems perfectly commonsensical to say that both overdeterminers are causes, and perfectly puzzling to say that neither are." Indeed, Lewis's later analyses both say that the individual overdeterminers are causes (Lewis, 1986, 2000).[4]

One might reply on behalf of Lewis's (1973a) analysis and the simple account like this: even though the individual rock throws of Suzy and Billy do not count as causes, the disjunction or mereological sum of both throws does. But this reply comes with costs and many open questions, as we argued in detail in Andreas and Günther (2025). To give a taste of the problems: What are disjunctive events? And where is their proper place

[4]These analyses overshoot in early and late preemption: the preempted would-be cause wrongly counts as a cause, respectively.

and time? If the disjunction or mereological sum of both throws is a cause of the window's shattering, is Suzy's throw alone also a cause? We think she could truthfully say in a court of law that her throw did not initiate a difference-making chain to the window's shattering. In this sense, she didn't cause it. And neither did Billy. But intuitively she did and so she should be held responsible for vandalising the window—and Billy as well. We will see that this problem accompanies counterfactual accounts to date. Our analysis, by contrast, has no such problems.

Another problem arises for the boulder scenario. A boulder is dislodged and rolls toward a hiker. The hiker sees the boulder coming and ducks, so that she does not get hit by the boulder. If the hiker had not ducked, however, the boulder would have hit her (Hitchcock, 2001, cf. p. 276). The structure of the boulder scenario can be represented by the following neuron diagram:

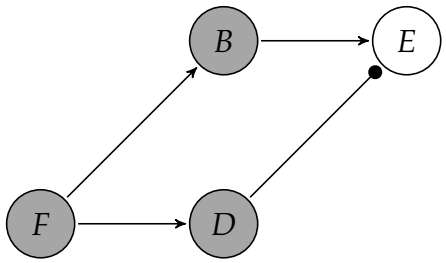

Figure 3: Short circuit

Hall (2007, p. 36) calls the network of Figure 3 a *short circuit*: the boulder's dislodgement (F) threatens to hit the hiker by a rolling boulder (B), and at the same time provokes an action—the ducking (D)—that prevents this threat from being effective ($\neg E$). The dislodgement of the boulder is not a cause of the hiker's remaining unscathed: F should not count as a cause of $\neg E$ because F creates *and* cancels the threat to bring about E (Paul and Hall, 2013, p. 216).

Indeed, the dislodged boulder makes no difference to the hiker's remaining unscathed. If the boulder had not been dislodged, the hiker would still have been unscathed. Lewis's analysis, somewhat surprisingly, misclassifies the dislodged boulder as a cause of the hiker's remaining un-

scathed. Barring backtracking, there is a difference-making chain: had the boulder not been dislodged, the hiker would not have ducked; and had the hiker not ducked, the boulder would have hit her. This means Lewis's solution to early preemption overshoots. There, Suzy's throw is a cause but no difference-maker. In the boulder scenario by contrast, the dislodged boulder is no cause but initiates a chain of difference-making. So Lewis's analysis is forced to count the dislodged boulder as a cause. To be clear, defining causation as the transitive closure of counterfactual dependence between actual events and absences solved early preemption. But the transitivity imposed on causation is of no help for late preemption and backfires in the boulder scenario. This result questions whether the deviation from counterfactual difference-making by imposing transitivity is warranted—or only motivated by solving early preemption. Many accounts in the counterfactual tradition no longer impose transitivity.

Lewis's analysis mainly fails for two reasons. The first concerns the very heart of counterfactual accounts: there is simply no counterfactual dependence in cases of redundant causation. And, second, repairing the absence of such a dependence by imposing transitivity backfires where causation is judged to be intransitive. Our epochetic theory, by contrast, is not susceptible to the problems. We neither need to rely on counterfactual dependence nor on imposing transitivity. Indeed, our theory solves overdetermination and preemption without further ado. And the solution to early and late preemption is analogous.

3 De Facto Dependence

We have seen that effects do not always counterfactually depend on their causes. But perhaps effects *do* always counterfactually depend on their causes when holding fixed certain actual events and absences. In early preemption, Suzy's throw is a genuine cause of the window's shattering but her throw does not make a difference to the shattering. Her throw *does* make such a difference, however, when holding fixed that Billy's rock does not touch upon the window. Perhaps Suzy's throw is a cause of the window's shattering because the shattering counterfactually depends on

her throw *given that Billy's rock has no impact on the window*. This is the idea of de facto dependence (Yablo, 2002).

The simple de facto account goes as follows:

> An event or absence C is a cause of a distinct event or absence E iff
>
> (i) C and E are actual, and
>
> (ii) there is a set F of "non-disjunctive" actual events and absences such that the counterfactual $(\neg C \wedge \bigwedge F) \mathrel{\square\!\!\rightarrow} \neg E$ is true, where $\bigwedge F$ denotes some conjunction of the members of F.

The de facto counterfactual $(\neg C \wedge \bigwedge F) \mathrel{\square\!\!\rightarrow} \neg E$ says "if C had not been actual but the events and absences in F had still been actual, E would not have been actual". The general idea is this: effects depend de facto on their causes—they counterfactually depend on their causes when the right surrounding events and absences are held fixed. The idea immediately poses the question: what are the "right events and absences" to be held fixed? While it seems clear in the preemption case, a general answer is difficult to give.

The simple de facto account does not restrict the set F of events and absences over and above imposing actuality on its members. As a result, counterfactual difference-making between actual events and absences is sufficient for causation. For $F = \emptyset$, the simple de facto account reduces to the simple counterfactual account. Hence, the simple de facto account—like Lewis's analysis—recognizes more causes than the simple counterfactual account.

The simple de facto account provides a straightforward and uniform solution to early and late preemption. However, the simple de facto account fails for the boulder scenario depicted in Figure 3—it identifies the dislodged boulder as a cause of the hiker's remaining unscathed. If the boulder had not been dislodged but it still had been rolling toward the hiker, the hiker would not have ducked and so would have been hit by the boulder. This de facto counterfactual is true. And yet, it seems strange. How

could the boulder not have been dislodged and still have rolled toward the hiker and hit her? This seems causally impossible. But wait. How could Billy's rock not have impacted upon the window if Suzy had not thrown? This seems causally impossible as well in the scenarios of early and late preemption. A defender of a de facto account should explain why we can hold fixed that Billy's rock does not touch the window in preemption, while we cannot hold fixed that the boulder is rolling toward the hiker.

Hitchcock (2001, pp. 297–8) thinks the answer is that we are not "willing to take seriously" certain far-fetched and contrary-to-fact combinations of events and absences. But why is holding fixed that Billy's rock does not touch the window if Suzy had not thrown a less far-fetched combination than the boulder rolling toward the hiker if it had not been dislodged? After all, both combinations involve counterfacts and violate the causal dependences between the events and absences. And no wonder. Accounts in terms of counterfactual dependence rely on counterfacts and "miracles" to explain causation. A dependence or law violation in point are true backtracking counterfactuals: if Suzy's rock had not touched the window, she would somehow still have thrown the rock—"with unfailing accuracy" we may add to the description of the preemption scenarios. Our analysis, by contrast, stays clear of these problems because it has no need for assuming causally impossible contrary-to-fact combinations of events and absences.

The simple de facto account succumbs to overdetermination. Suppose C and A overdetermine E. Then there is no set F of "non-disjunctive" actual events and absences such that $(\neg C \wedge \bigwedge F) \boxdot\!\!\rightarrow \neg E$ is true. And similarly, A is not a cause of E. Yablo (2002) writes about overdetermination:

> But then what does cause the window to break? Not the conjunction of the two throws, since the effect could too easily have occurred without it. Not the disjunction, because we are hard put to regard the disjunction as a genuine event. Could it be that nothing causes the window to break? This goes somewhat against the grain. An event that was caused (the breaking was not a miracle!) should, one feels, have causes. (p. 139)

Yablo agrees with "Lewis that the case is *intuitively* undecidable" and "can be left as 'spoils to the victor'." (ibid.) His own de facto account aims to

capture our putative indecision as to whether the individual overdeterminers count as causes.

Yablo (2002, 2004) proposes a more sophisticated de facto account:

> An actual event or absence C is a cause of a distinct one E iff
>
> (1) there is some set F of actual events and absences such that E counterfactually depends on C when the events and absences in F are held fixed, and
>
> (2) the set F is *right* and *more natural* than any *wrong* alternative.

Indeed, Yablo defines that C de facto depends on E if (1) and (2) are satisfied. To understand (2), we need to explain what a right and a wrong set F of actual events and absences is, and what it means that some F is more natural than another.

The distinction between right and wrong events and absences to be held fixed is relative to candidate cause and effect. Fix candidate cause C and putative effect E. The idea is then to balance the set of *needs* the putative effect E has, holding F fixed, with the set of needs it would have had, had the candidate cause C not occurred. F is, intuitively speaking, right if E has fewer needs holding F fixed as compared to the needs it would have had if C had not occurred.

We may, roughly, express the set of needs the effect E would have had, had the candidate cause C not occurred by a set S_C of events and absences on which E would have depended if the candidate cause C had not been actual:

$$S_C = \{A \mid (\neg C \wedge \neg A) \,\Box\!\!\rightarrow \neg E\}.$$

The natural language phrasing suggests that the counterfactual defining S_C is $\neg C \,\Box\!\!\rightarrow (\neg A \,\Box\!\!\rightarrow \neg E)$: had C not been actual, E would have depended on A. Yablo (2002, p. 136, fn. 17) deviates from Lewis's semantics in assuming that the two counterfactuals are equivalent, at least for the purposes of his theory.

The set of needs an effect E has, holding F fixed, may roughly be expressed by a set P_F of actual events and absences on which E would have

depended if the events and absences in F are held fixed:

$$P_F = \{D \mid (\bigwedge F \wedge \neg D) \,\Box\!\!\rightarrow \neg E\}.$$

The set F of actual events and absences to be held fixed is wrong if S_C is a non-empty subset of P_F; otherwise F is right.

Here is an immediate corollary. Suppose E counterfactually depends on C outright—without holding fixed any set F of actual events and absences. Hence, had C not been actual, E would not have been actual. But then, had C not been actual, E would have no needs: there are no events and absences E would depend on—for there is no E that could depend on anything. Hence, $S_C = \emptyset$ if E counterfactually depends on C outright. And so any F is right, in particular the empty or "tautological" F. C is then a cause of E.

Let us illustrate the distinction between right and wrong sets F of what can be held fixed by revisiting late preemption. If Suzy had not thrown her rock ($\neg C$), the breaking of the window (E) would have counterfactually depended on Billy's throw (A). Hence, $A \in S_C$. The intuitively right absence to be held fixed is that Billy's rock does not touch upon the window: $F = \{\neg B\}$. When holding this F fixed the window's shattering E does not counterfactually depend on Billy's throw A. Hence, $A \notin P_F$. And so the non-empty S_C is not a subset of P_F, which means that $F = \{\neg B\}$ is right. Holding F fixed reduces the needs E has and so reveals a counterfactual dependence of E on C.

If we suppose that $F = \{\neg B\}$ is more natural than any wrong alternative, Suzy's throw is a cause of the window's shattering. When Suzy had not thrown her rock while holding F fixed, the window would not have shattered. But what does "more natural" mean? Some events and absences F to be held fixed are gerrymandered or ad hoc. The mereological sum or "disjunction" $C \vee \neg E$, for example, seems far from natural. And indeed, holding this "actual" disjunction fixed—C occurs—E counterfactually depends on C in general.

Fortunately for Yablo, we have some intuitions what sets F of actual events and absences are more natural than others—and the disjunction $C \vee \neg E$ is

rather less natural. The problem is that we need to rely on our intuitions about naturalness. For Yablo (2002, p. 133, fn. 11) proposes to

> postpone (= ignore) the question of what is the best thing to mean by "natural". This is partly because I am uncertain about it, in particular about the extent to which cognitive and cultural factors are allowed to come in.

The boulder scenario may illustrate the problem of deciding what F counts as natural or more natural. A dislodged boulder (F') causes the ducking of a hiker (D), which in turn causes the hiker to remain unscathed ($\neg E$). But it is counterintuitive to say that the dislodging of the boulder causes the hiker to remain unscathed.

Is the boulder's dislodgement a cause of the hiker's remaining unscathed on Yablo's account? It depends. The hiker's remaining untouched by the boulder counterfactually depends on the dislodgement of the boulder when holding fixed that the boulder is rolling towards the hiker. Hence, the boulder's dislodgement counts as a cause of the hiker's remaining unscathed if $F = \{B\}$ is right and more natural than any wrong alternative. $F = \{B\}$ is right because the corresponding set $S_{F'}$ of events and absences on which $\neg E$ would have depended on if F' had not occurred contains $\neg B$ and so is no subset of P_F. But is $F = \{B\}$ more natural than any wrong alternative?

Holding fixed that the boulder is rolling towards the hiker when counterfactually assuming that the boulder has not been dislodged seems somewhat unnatural. However, it seems to be on a par with assuming in late preemption that Billy's rock does not touch upon the window when counterfactually assuming that Suzy does not throw her rock. For then Billy still throws his rock but it somehow fails to touch the window. Yablo's account can perhaps provide the desired result that the dislodgement of the boulder is not a cause of the hiker's remaining unscathed, but it needs to say more about the comparative naturalness of the sets F of actual events and absences. Without such an amendment, Yablo's account does not deliver clear verdicts on various causal scenarios.

Let us revisit overdetermination. Suzy and Billy each throw a rock at a

window and the window shatters. Neither Suzy's throw (C) nor Billy's throw (A) alone make a difference to the window's shattering (E). On Yablo's account, Suzy's throw is a cause of the window's shattering iff there is some F of actual events and absences such that (1) the counterfactual $(\neg C \wedge \bigwedge F) \boxdot\!\!\rightarrow \neg E$ is true, and (2) F is right and more natural than any wrong alternative.

For determining whether C is a cause of E, Yablo (2002, p. 140) *assumes* that the following two sets of actual events and absences are most natural:

$\{\neg A \vee C\}$: Billy's rock does not hit the window without Suzy's.

$\{A \vee \neg C\}$: Suzy's rock does not hit the window without Billy's.

"Disjunctive events" usually count as less natural—if not unnatural outright. But Yablo's account requires of the right F not that it is natural—only that it is *more natural* as compared to the wrong alternatives.

No rock would hit the window if Suzy hadn't thrown hers holding fixed the disjunction $\neg A \vee C$. This means the window's shattering counterfactually depends on Suzy's throw holding fixed that Billy's throw does not hit the window or Suzy's does: $\neg C \wedge (\neg A \vee C) \boxdot\!\!\rightarrow \neg E$. Now, $\{\neg A \vee C\}$ is right because $S_C = \{A \mid (\neg A \wedge \neg C) \boxdot\!\!\rightarrow \neg E\}$ is not empty and not a subset of $P_{\neg A \vee C} = \{D \mid ((\neg A \vee C) \wedge \neg D) \boxdot\!\!\rightarrow \neg E\}$. A is a member in S_C but not in $P_{\neg A \vee C}$.

By symmetrical reasoning, $\neg A \wedge (A \vee \neg C) \boxdot\!\!\rightarrow \neg E$ is true. And $\{A \vee \neg C\}$ is right because $S_A = \{C \mid (\neg A \wedge \neg C) \boxdot\!\!\rightarrow \neg E\}$ is not empty and not a subset of $P_{A \vee \neg C} = \{D \mid ((A \vee \neg C) \wedge \neg D) \boxdot\!\!\rightarrow \neg E\}$. C is a member in S_A but not in $P_{A \vee \neg C}$. So far so good.

Is Suzy's throw a cause after all? Well, this depends on whether there is a more natural but wrong set of events and absences such that holding them fixed the window's shattering counterfactually depends on Suzy's throw. We have just seen a candidate for such a more natural but wrong set: $\{A \vee \neg C\}$. The window's shattering would depend on Billy's throw holding fixed $A \vee \neg C$ that Suzy's rock does not hit the window or Billy's does: $A \in P_{A \vee \neg C}$. Indeed, S_C is a non-empty subset of $P_{A \vee \neg C}$, which makes $\{A \vee \neg C\}$ wrong. The question is now whether $\{A \vee \neg C\}$ is more or less

natural than $\{\neg A \vee C\}$. As noted above, Yablo assumes that they are both the most natural sets of events and absences for determining whether C is a cause of E. By symmetrical reasoning, both singleton sets are the most natural sets for determining whether A is a cause of E.

Suzy's throw is a cause of the window's breaking iff "more natural" means *at least as natural*. Indeed, then the two throws individually count as causes. If "more natural", by contrast, means *strictly more natural*, then neither throw counts as a cause. Yablo (2002, p. 140) thinks the different readings of "more natural" explain the putative indecision whether the individual throws count as a cause. This is an elegant explanation provided you are undecided. If you are decided that the individual overdeterminers are causes, one must adopt the weaker reading.

Let us return to late preemption. We show now that Billy's throw—the preempted would-be cause—may on Yablo's account come out as a cause of the window's shattering after all (Paul and Hall, 2013, cf. p. 115). For this to be seen, consider the set F of actual events and absences which says that "Suzy's rock is never far ahead of Billy's rock." Yablo imposes no restrictions on the candidate sets of actual events and absences. Hence, F seems to be a fair choice. If Billy had not thrown while holding fixed that Suzy's rock is never far ahead of Billy's, the window would not have shattered. For Suzy's rock would have remained close to Billy and the rock in his hands—it would not have come close to the window.

It remains to show that F is right and more natural—or at least as natural as—any wrong alternative. Suppose Billy had not thrown his rock. The window's shattering would then have depended on Suzy's throw ($C \in S_A$). However, holding fixed that Suzy's rock is never far ahead of Billy's, the window would still shatter if Suzy had not thrown her rock ($C \notin P_F$). Holding fixed F is compatible with Billy's rock being far ahead of Suzy's. The set S_A of events the window's shattering would have depended on in the absence of Billy's throw is not empty and no subset of the set P_F of events the shattering depends on, holding F fixed. Hence, F is right.

Is there some wrong F' (strictly) more natural than F which undermines the status of Billy's throw as a cause? Perhaps, but it is hard to see this wrong set of actual event and absences. This is, in part, because F is not

that unnatural: in the actual scenario, Suzy's rock is never far ahead of Billy's rock. Without further elaboration on which sets to be held fixed are more natural than others, it seems that Billy's throw is mistakenly counted as a cause in late preemption. This is, of course, bad news for Yablo's account.

We have seen that Yablo proposes an elaborate de facto account. The basic strategy is to reveal hidden counterfactual dependences by holding the right events and absences fixed. This promising strategy of de facto accounts is coupled with a notion of comparative naturalness. We have seen that this notion, which Yablo did not work out in sufficient detail, infects his account with imprecision: the verdicts for causal scenarios like the boulder scenario, symmetric overdetermination, and late preemption depend on which right sets of actual events and absences are judged to be more natural than its wrong alternatives. Yablo does not say enough to properly regiment our judgements about comparative naturalness. This does not mean that his de facto account is beyond repair. But as it stands, it only provides clear verdicts when our judgments of comparative naturalness are beyond doubt—not so often.

There are de facto accounts of causation which—unlike Yablo's—have no need for judgments of comparative naturalness. They provide clear verdicts by relying on an interventionist semantics of conditionals. We turn to these causal model accounts next.

4 Counterfactual Interventions

Pearl's (2000) framework of causal models opened up novel ways to define causation in terms of counterfactuals. The causal models introduced by Andreas and Günther (forthcoming) deviate from the standard account by being based on classical propositional logic. The deviation was not made for its own sake, but rather enabled us to study inferential pathways in the first place. For ease of notation and simplicity, we continue to use our format when explaining counterfactual approaches to causation with causal models. To this end, we need to generalize the account of causal models from Andreas and Günther (forthcoming, Ch. 2, Sec. 1) in two ways.

First, we need to understand interventionist conditionals for counterfactual antecedents. So far, we have explained interventions on a causal model $\langle M, V\rangle$ only for literals and sets of literals which are consistent with V. However, we can explain counterfactual interventions in terms of interventions which result in a consistent model $\langle M_I, V' \cup I\rangle$. Recall that I stands for the set of literals by which we intervene on $\langle M, V\rangle$.

Suppose we want to intervene on $\langle M, V\rangle$ by a set U of literals, where U is not consistent with the union of M and V. Now, let V_{N-U} be the subset of V which contains all literals in V such that the variable of the literal satisfies two conditions: it is a non-descendant of, and different from, all variables which have a literal in U. In formal terms, $L_A \in V_{N-U}$ iff $L_A \in V$ and A is a non-descendant of, and different from, all variables B for which there is $L_B \in U$. Then we can define a counterfactual interventionist conditional as follows:

$$\langle M, V\rangle[U] \models \phi \text{ iff } \langle M, V_{N-U}\rangle[U] \models \phi.$$

Recall that $\langle M, V_{N-U}\rangle[U]$ is defined as the model $\langle M_U, V_{N-U} \cup U\rangle$. The latter is a consistent model since M_U does not contain the structural equations of variables which have a literal in U. And V_{N-U} does not contain any information about variables on which we intervene by U. Nor does V_{N-U} contain any information on the descendants of such variables.

The second generalization concerns the arity of variables in a causal model $\langle M, V\rangle$. So far, we have been working with binary variables. The main reason for this restriction is simplicity. Binary variables suffice to capture virtually all scenarios of actual causation which have received significant attention in literature. The counterfactual accounts of causation using causal models, however, cover non-binary variables. A comparison to them should therefore explain how our formalism can be extended to non-binary variables.

Here is a simple example of a non-binary variable. Suppose Suzy can throw a rock at a bottle, at a window, or not throw a rock at all. These possibilities may well be expressed by propositional variables. However, let's suppose we want to express them by a ternary variable X. Obviously, X can assume three values. Let's take b for a rock throw at a bottle, w

for one at a window, and n for no rock throw at all. The different value assignments can now be expressed by the sentences $X = b$, $X = w$, and $X = n$.

Now, this notation can easily be translated into atomic sentences in first-order logic, the latter having the benefit of being more explicit. Suppose s stands for Suzy, and t is a first-order function, which stands for throwing a rock. Then the three different value assignments of the variable X can be expressed by the following atomic sentences: $t(s) = b$, $t(s) = w$, and $t(s) = n$. Note that the negations of these atomic sentences are well formed sentences of first-order logic too. So we have positive and negative literals as we do in propositional logic.

Just one qualification is needed to make the translation of value assignments into atomic sentences consistent with the framework of first-order logic: we need to restrict the domain and the range of the first-order functions used to express value assignments of non-binary variables. Such a restriction is more than intuitive. Most expressions of functions in natural language and mathematics are not understood with an unrestricted domain of interpretation. The age of an object is a simple example. For abstract objects, it simply does not make sense to ask how old they are. We can restrict the domain and the range of a function in many-sorted first-order logic. This type of first-order logic is often tacitly assumed when functions are used to represent some piece of knowledge and reasoning. The details of our generalized causal models are explained in an appendix on the logic of causal models (Andreas and Günther, forthcoming, App. A).

5 De Counterfacto Dependence in Causal Models

Let us begin with the above scenario of overdetermination: Suzy and Billy throw a rock at a window. Their rocks hit the window at the same time, and the window shatters. Notably, each rock throw is sufficient for the window to shatter. For the overdetermination scenario, the valuation V

may be given by the set $\{C, A, E\}$, which says that Suzy throws, Billy throws, and the window shatters. The set M of structural equations just contains $E = C \vee A$, which says that the window breaks iff Suzy or Billy throws. A causal model account makes explicit what events and absences, and relations between those we are to consider.

Interventionist counterfactuals are suitable to spell out de facto and *de counterfacto* dependence. Relative to the causal model of overdetermination, for example, Suzy's throw is a cause of the window's shattering because the window's shattering counterfactually depends on Suzy's throw if we hold *the counterfact that Billy does not throw* fixed by intervention. Holding fixed the variable A at its counterfactual value $\neg A$ reveals a hidden counterfactual dependence of the effect E on its cause C. This is a straightforward solution to the problem of overdetermination.

The causal model accounts can solve overdetermination if they lift the restriction of *de facto* accounts that only actual events and absences can be held fixed. But this move from de facto to de counterfacto dependence opens the door for a plethora of new problems. Billy's throw (A), for example, comes out as a cause of the window's shattering (E) in early preemption on the simple de counterfacto account: A and E is actual and the de counterfacto conditional $(\neg A \wedge \bigwedge CF) \,\Box\!\!\rightarrow \neg E$ is true for the set $CF = \{\neg C, \neg D\}$ of facts and counterfacts.[5] What is direly needed is a restriction on which facts and counterfacts can be held fixed. And in the best case this restriction should be clear and well-motivated.

6 Active Routes

The de counterfacto account by Hitchcock (2001) centres on the notion of *active route*, which resembles our notion of active path. In fact, we took some inspiration from Hitchcock when working out the latter notion. Our notion of active path may be seen as the factual counterpart to Hitchcock's notion of active route. Specifically, we have aimed to reconstruct some

[5] Another among the many problems for the simple de counterfacto account is the scenario discussed by Paul and Hall (2013, pp. 198–9).

concept of *factual dependence* of the effect on the candidate cause.

Hitchcock (2001) says causation is tantamount to the existence of an active route from cause to effect in a causal model. Such an active route effectively requires that a cause makes a difference to its effect along some route when holding fixed some surrounding events and absences. The set of facts and counterfacts, which may be held fixed, is thus restricted by the constraint that the active route from cause to effect must remain intact. This is a clear and well-motivated restriction on the face of it. So let us consider Hitchcock's de facto proposal in more detail:

> The value assignment C is a cause of the value assignment E relative to a causal model $\langle M, V\rangle$ iff there is an *active route* from the variable C to the variable E in $\langle M, V\rangle$.

A route—or directed path in our terminology—between the two variables C and E of a causal model is a sequence of variables $\langle C, D_1, ..., D_n, E\rangle$, where each variable in the sequence is on the right-hand side of the structural equation of its successor in the sequence. A route $\langle C, D_1, ..., D_n, E\rangle$ is *active* in a causal model $\langle M, V\rangle$ iff $\neg C \,\Box\!\!\rightarrow \neg E$ is true in the causal model $\langle M', V\rangle$, where M' is obtained from M as follows: if the variable D is on some route between C and E, but does not belong to the route $\langle C, D_1, ..., D_n, E\rangle$, then remove the equation for the variable D and set it to its actual value in V. For C to be a cause of E, Hitchcock effectively requires that E counterfactually depends on C when the variables between C and E which are *not* on a specific route from C to E are held fixed at their actual values. He offers thereby a de facto account of causation which—unlike Yablo's—leaves no doubt what events and absences may be held fixed.

Hitchcock's account can solve early and late preemption alike, depicted by Figure 1 and 2. In both scenarios, there is an active route from C over D to E when holding the variable B fixed at its actual value $\neg B$. Hence, C is a cause of E. By contrast, there is no active causal route from A over B to E. There is exactly one directed path from A to E—$\langle A, B, E\rangle$—and it is not active: E does not counterfactually depend on A when holding other variables fixed at their actual values. Hence, A is not a cause of E. This is a clear and promising result.

What about overdetermination? Well, there is neither an active route from Suzy's throw (C) to the window's shattering (E) nor from Billy's (A). If C had not occurred, E would still have occurred. And symmetrically for A. To solve the problem of overdetermining causes, Hitchcock weakens the notion of active route. The idea is that the values of those variables, which do not lie on the considered route between C and E, may be changed as long as these changes do not change the values of the variables on the considered route. In overdetermination, the value of the variable A can be changed to $\neg A$ without affecting the values of C and E. If A had taken the value $\neg A$, C and E would still have taken the values C and E, respectively. So we can keep A fixed at its non-actual value $\neg A$ to reveal that E would have been absent if C had been absent. Hitchcock argues that holding fixed A at its non-actual value $\neg A$ reveals a counterfactual dependence of E on C that is hidden in the actual situation. And this hidden counterfactual dependence is taken to be sufficient for causation. Hitchcock's weakly active route thus generalizes his de facto account to a de counterfacto account.[6]

More precisely, a route $\langle C, D_1, ..., D_n, E\rangle$ is *weakly active* in $\langle M, V\rangle$ iff there is a possibly empty set $\mathcal{W}$ of variables of the considered causal model all of which (a) do not lie on $\langle C, D_1, ..., D_n, E\rangle$ and (b) $\neg C \,\Box\!\!\rightarrow \neg E$ is true in the causal model $\langle M', V'\rangle$, where M' is obtained from M as follows: for each W in $\mathcal{W}$, remove the equation for the variable W and set it to a value which does not change the value of any variable lying on $\langle D_1, ..., D_n, E\rangle$; V' is the resulting value assignment to the variables. Note that an active route is also weakly active. Moreover, if $\neg C \,\Box\!\!\rightarrow \neg E$ is true in the causal model $\langle M, V\rangle$, then there is an active route from C to E. Counterfactual dependence in a causal model is thus sufficient for causation on Hitchkock's account.

In the overdetermination scenario, the route $\langle C, E\rangle$ is weakly active because there is $\mathcal{W} = \{A\}$ and setting A to its non-actual value $\neg A$ does not affect the values of C and E; and yet setting the variable A to its non-

[6]It should be noted that Hitchcock's (2001) active and weakly active route essentially correspond to the natural and causal beam of Pearl (2000, Ch. 10), respectively. Moreover, the account of actual causation provided by Woodward (2003, p. 84) corresponds to the notion of weakly active route and so to the notion of causal beam as well.

actual value $\neg A$ reveals the counterfactual dependence of E on C. Hence, C is a cause of E on Hitchcock's liberalised account. And by symmetrical reasoning A is a cause of E. So far so good.

But trouble is not far to seek. Recall the boulder scenario depicted in Figure 3. The boulder's dislodgement (F) threatens to hit the hiker by a rolling boulder (B), and at the same time provokes an action—the ducking (D)—that prevents this threat from being effective ($\neg E$). The boulder's dislodgement is not a cause of the hiker's remaining unscathed. However, Hitchcock's account says so. For this to be seen, let's consider the causal model corresponding to the neuron diagram in Figure 2. Observe that the variable B is between the variables F and E, but does not lie on the route from F over D to E. Holding B fixed at its actual value B, $\neg E$ counterfactually depends on F. Hence, the route $\langle F, D, E\rangle$ is active and so F wrongly counts as a cause of $\neg E$ on Hitchcock's account—and hence also on his liberalized account.

The decisive de facto counterfactual is, again, this: if the boulder had not been dislodged ($\neg F$) but still had rolled towards the hiker (B), the hiker would have been hit by the boulder (E). Recall from our discussion of Yablo's account that it may seem unnatural to hold fixed that the boulder is rolling towards the hiker when counterfactually assuming that the boulder has not been dislodged. Hitchcock (2001, p. 297) expresses a similar sentiment about the counterfactual:

> the relevant piece of counterfactual reasoning would go as follows: suppose that the boulder had been present at a point one metre from Hiker's head and flying toward him, and suppose moreover that it had never fallen in the first place. Since it never fell, Hiker would not have seen it coming and would not have ducked; since it would have been there, one metre from his exposed head, it would have hit him and he would not have survived. This counterfactual reasoning is correct, but bizarre. If the boulder never fell, how did it get to be there, one metre from Hiker's head?

The question is fair enough. And so is this one about the preemption cases: if Suzy had never thrown her rock, how could Billy's rock still not touch

the window? A de facto account of causation should explain why the former counterfactual reasoning is bizarre while the latter is allegedly not. For counterfactual accounts trade in non-actual possibilities and miracles to explain causation. Assuming that this is a valid strategy, the question is why some non-actual possibilities and violations of law are fine while others are not. Our analysis of causation, by contrast, has no need for such counterfactuals—be they bizarre or not.

7 Sophisticated Causal Model Accounts

Halpern and Pearl have proposed several de facto and de counterfacto accounts of causation that rely on causal models. Halpern (2016, Ch. 2) distinguishes between three accounts of causation: the original Halpern-Pearl definition in Halpern and Pearl (2001), the updated Halpern-Pearl definition in Halpern and Pearl (2005), and the modified Halpern-Pearl definition in Halpern (2015). The updated definition supersedes the original one. We will therefore confine our discussion to the latter two definitions.

The template for all Halpern-Pearl accounts of causation is as follows. Where X is a set of value assignments and ϕ a propositional formula,

> X is a cause of ϕ relative to the causal model $\langle M, V\rangle$ iff all of the following conditions are satisfied:
>
> (HP1) $\langle M, V\rangle \models \bigwedge X \wedge \phi$, where $\bigwedge X$ is some conjunction of all the members of X.
>
> (HP2) To be filled in below.
>
> (HP3) There is no proper subset X' of X such that (HP1) and (HP2) are satisfied.

Halpern and Pearl consider not only single value assignments as candidates for causes and effects. They allow for sets of variable assignments to be causes and for propositional formulas to be effects. (HP1) says that both cause and effect must be actual. (HP3) ensures that any cause is minimal

in this sense: X contains no elements that are unnecessary for satisfying (HP1) and (HP2).

The modified Halpern-Pearl definition is the newest and, perhaps surprisingly, simplest of their accounts. The idea is reminiscent of the one behind the simple de facto account: test for counterfactual dependence when certain variables are held fixed by intervention at their actual values. Here is the condition making the modified definition a de facto account:

(HP2^m) There is a set $W \subset V$ of actual value assignments such that $\langle M, V\rangle[W][\neg X] \models \neg\phi$, where the bold $\neg$ is elementwise negation.

For X to be a cause of ϕ relative to $\langle M, V\rangle$, (HP2^m) requires that ϕ counterfactually depends on X when holding the variables appearing in W fixed at their actual values. When no variables are held fixed $W = \emptyset$, the condition reduces to counterfactual dependence of ϕ on X. Outright counterfactual dependence is thus sufficient for causation provided that both X and ϕ are actual (HP1) and X is minimal (HP3). Recall that value assignments may be expressed by sets of literals in our account of causal models. On this understanding, V, W, and X are sets of literals.

Like the simple de facto account, the modified definition solves early and late preemption. Hold fixed that Billy's rock does not touch upon the window $W = \{\neg B\}$. If Suzy then had not thrown her rock, the window would not have shattered. Suzy's throw is thus a cause of the window's shattering. By contrast, Billy's throw is not. Holding fixed any variables at their actual values, had Billy not thrown, the window would have shattered anyways.

(HP2^m) can be seen as a liberalization of Hitchcock's active route. There is no restriction to keep only variables fixed which do not lie on a specific directed path between cause and effect. Indeed, the modified definition does not explicitly mention any directed path at all. And yet, there must be a directed path from a genuine cause X to its effect ϕ. For ϕ only counterfactually depends on X while holding some variables at their actual values if there are directed paths from variables appearing in X to variables appearing in ϕ. There is no counterfactual dependence when all variables

are held fixed at their actual values. There is likewise no counterfactual dependence when the variables on each directed path between cause and effect are held fixed. Hence, whenever the modified definition claims causation, there must be at least one directed path from a cause variable to the effect such that the variables on this path are not held fixed at their actual values.

Like the simple de facto account and Hitchcock's active route, the modified definition allows only to hold fixed variables at their actual values. As a consequence, individual overdeterminers do not count as causes. There is no set of actual value assignments such that $\neg E$ would be the case if $\neg C$ were the case in the causal model of overdetermination. Hence, the overdeterminer C does not count as a cause of E. This being said, the set $\{C, A\}$ of variable assignments counts as a cause of E. Holding nothing fixed, if $\neg C$ and $\neg A$ were the case, $\neg E$ would be the case. While the only cause of the effect is the set, its members are *parts* of the cause. And parts of causes are "what we think of as causes"—or so says Halpern (2016, p. 25). Perhaps a cause is nothing but an element of a minimal set which makes a counterfactual difference (Andreas and Günther, 2021). This move would pose the mereological problems we mentioned above and would require a suitable notion of minimality. It is up to the defenders of counterfactual difference-making to explore this avenue.

One must wonder, however, why an overdeterminer is counted as part of a cause while the set $\{C, A\}$ is not counted as a cause in a conjunctive scenario, where the occurrence of both events C and A is necessary and sufficient for the effect to occur. As Andreas and Günther (2021) pointed out, the set $\{C, A\}$ is intuitively *the*—or at least a—cause of the effect. Without revisiting the metaphysical problems posed by parts of causes, we note this slight tension of the modified definition with our common sense judgments. We will revisit this issue when discussing Gallow's (2021) theory of causation in Section 10 below.

The updated Halpern-Pearl definition counts individual overdeterminers as causes. The underlying reason is that the updated definition is a de counterfacto account—it allows to hold fixed certain variables at non-actual values. On the downside, the updated definition is more involved than the modified one. The condition (HP2^u) is split into two parts, (a)

and (b). Part (a) tests for counterfactual dependence of the candidate effect ϕ on the candidate cause X when certain variables are held fixed by intervention. Unlike (HP2^m), part (a) allows to hold fixed certain variables at non-actual values: it allows for non-actual *contingencies*—Halpern and Pearl's name for value settings of variables by interventions. Part (b) constrains the choice of contingencies. The idea is, roughly speaking, that the effect ϕ must still be actual under the contingency chosen in part (a) if the remaining variables are set to their actual values.

A statement of the second condition of the updated definition requires us to make the distinction between a set of variables and their value assignments explicit. Let $\mathcal{V}$ be the set of variables that appear in the value assignment V, and $\mathcal{X}$ be the set of variables appearing in the value assignment X. We can now state the second condition as follows:

(HP2^u) There is a partition of the set $\mathcal{V}$ of variables into sets $\mathcal{W}$ and $\mathcal{Z}$, where $\mathcal{X} \subseteq \mathcal{Z}$, such that

(a) there is a possibly non-actual value assignment W to the variables in $\mathcal{W}$ and a value assignment $\neg X$ to the variables in $\mathcal{X}$ such that $\langle M, V\rangle[W][\neg X] \models \neg\phi$, and

(b) for all subsets $\mathcal{W}'$ of $\mathcal{W}$ and all subsets $\mathcal{Z}'$ of $\mathcal{Z}$, $\langle M, V\rangle[W'][Z'][X] \models \phi$, where W' is the value assignment that corresponds to W of (a) restricted to the variables in $\mathcal{W}'$, and Z' is the value assignment that corresponds to the actual values of the variables in $\mathcal{Z}$ restricted to $\mathcal{Z}'$.

Part (a) says that ϕ counterfactually depends on X under the possibly non-actual contingency W. Part (b) says, roughly, that ϕ is still actual under the contingency W if the variables in $\mathcal{Z}$ are set to their actual values. Part (b) so understood resembles Hitchcock's weakly active route: the possibly non-actual contingency W cannot change the value of certain other variables. Indeed, Halpern and Hitchcock (2010, p. 392) suggest to conceive of the variables of Z as making up the 'causal path' from the candidate cause to its effect. However, the subclause says more: ϕ is still actual even if we intervene by any subset W' of the value assignments W and any subset Z'

of the actual assignments $Z \subset V$ alongside X. This essentially means that a cause X must be *sufficient* for its effect ϕ under certain actual and non-actual contingencies (Halpern and Pearl, 2005, cf. p. 854). While part (a) is a counterfactual condition for the cause candidate X, part (b) is not. The latter is rather a sufficiency or production condition for X, as Halpern and Pearl (2005, p. 867) acknowledge. In this sense, their updated definition is not a purely counterfactual account of causation.

How does the updated definition solve overdetermination? Well, there is a non-actual value assignment $\neg A$ such that $\neg E$ would be the case if $\neg C$ were the case under this non-actual value assignment. Part (a) is satisfied. Part (b) is also satisfied. To see this, note that setting C by intervention is *sufficient* for E to be actual, and no combination of settings of any subsets W' and Z' undoes this *sufficiency*.

The modified and updated definitions succumb to the boulder scenario for reasons similar to Hitchcock's account. Hold fixed that the boulder is rolling towards the hiker. The dislodged boulder then makes a difference to the hiker's remaining unscathed. This points to a principled problem of de facto and de counterfacto accounts (Andreas and Günther, forthcoming). The general strategy to test for counterfactual dependence while holding certain variables fixed by intervention at certain values allows to solve preemption and other problematic scenarios. However, the same strategy systematically delivers the wrong results in other scenarios such as the boulder scenario and the simple switch, as we will see in the next section. And the strategy backfires in the latter scenarios for the same reason it succeeds in the former. Our analysis, by contrast, is not susceptible to this principled problem.

It should be noted that the more recent counterfactual theories of Gallow (2021) and Andreas and Günther (2021) are likewise not susceptible to the principled problem. Indeed, both theories solve the set of scenarios that troubles causal model accounts which rely on de facto and de counterfacto dependence. The analysis by Andreas and Günther (2021) relies on a removal of information similar to the suspension of judgment of our analysis. We will compare our analysis to Gallow's (2021) theory in Section 10 below.

8 Switches

Switching scenarios mean trouble for most counterfactual accounts. Consider this simple switching scenario: Flipper is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch (F), so that the train travels down the right track (R), instead of the left ($\neg L$). Since the tracks reconverge up ahead, the train arrives at its destination all the same (E). The commonsensical judgment is that flipping the switch is not a cause of the train's arrival. Flipping the switch makes no difference to the train's arrival: *the train arrives at its destination all the same* independent of the flipping. Flipping the switch is, however, a cause of the train's travelling on the right track, and the train's travelling on the right track is a cause of the train's arrival (Paul and Hall, 2013, p. 232). This is yet another example suggesting that causation is not transitive. The switching scenario may be represented by a simple dependency diagram:[7]

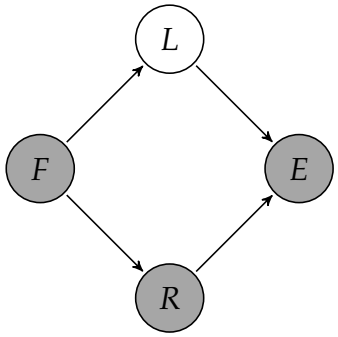

Figure 4: Simple switch

It seems that counterfactual accounts should have no problem with the simple switch. The flipping of the switch does not make a difference

[7]We think it is hard to represent switching scenarios by neuron diagrams. One reason is that the two positions of a switch are assumed to be symmetric, while the firing and the non-firing of a neuron are not. A further reason is that a firing "switch neuron" would activate a neuron and inhibit another. But this is too much: a switch should only determine the path by which an event is brought about. Neuron diagrams introduce an asymmetry with respect to the position of a switch, while there should be none.

to the train's arrival. By contrast, the flipping makes a difference to the train's travelling on the right tracks. The simple counterfactual account gets the verdicts right because it stays true to the guiding idea of counterfactual dependence between actual events: causation is identified with difference-making. We have seen that Lewis (1973a) deviates from this idea: he thinks of causes as initiators of difference-making chains. Hence, his analysis misclassifies the flipping as a cause of the train's arrival. The flipping of the switch makes a difference to the train's travelling on the right tracks. Barring backtracking, the train's travelling on the right tracks makes a difference to the train's arrival. By the transitive closure imposed on the one-step counterfactual dependences, Lewis (1973a) is forced to say that the flipping is a cause of the arrival. And his analysis is forced to say so in the basic, the realistic, and the realistic basic switch.[8]

Lewis's solution of early preemption backfires in the simple and other switching scenarios. For the former, his analysis correctly claims causation without difference-making. However, the flipping makes no difference to the train's arrival and is not a cause thereof. And yet, Lewis's analysis claims causation. This illustrates two points. First, an initiator of a difference-making chain does not always make a difference to the endpoint of that chain. Second, an initiator of a difference-making chain is not always a cause of the chain's endpoint.

The de facto accounts do not impose transitivity on causation. And yet they wrongly count the flipping as a cause of the train's arrival in the simple, the basic, the realistic, and the realistic basic switch (see Andreas and Günther (forthcoming, Chs. 4&6)). The reason is in each switching scenario that E counterfactually depends on F when holding $\neg L$ fixed. Hence, the simple de facto account, and the de facto accounts due to Hitchcock (2001) and Halpern (2015), and the de counterfacto account of Halpern and Pearl (2005) succumb to this verdict. There is an active route from F over R to E: holding the off-path variable L fixed at its actual value induces a counterfactual dependence of E on F. This points again to the principled problem of de facto and de counterfacto accounts. Allowing

[8]Lewis (2000) still imposes transitivity on his analysis of causation as influence. As a consequence, his refined analysis is also forced to say that the flipping of the switch is a cause of the train's arrival in the switch scenarios.

for actual contingencies, to use Halpern and Pearl's term, solved preemption, but leads to trouble in switching scenarios, the boulder scenario, and others such as extended double prevention.

The verdict in the respective switching scenarios as to whether or not the flipping is a cause of the train's arrival is not so clear cut on Yablo's (2002) de facto account. F is a cause of E iff the set $F' = \{\neg L\}$ of actual events and absences is more—or at least as—natural than any wrong alternative. Yablo (2004, p. 135) argues that the empty $F' = \emptyset$ is more natural than $F' = \{\neg L\}$. This may well be true. Even though you may as well have the opposite intuition that the empty set is no more natural, as Yablo (2002, pp. 145n) seems to acknowledge. The opposite intuition is the intuition Yablo relied on in early preemption: holding the empty set fixed is no more natural than holding fixed that Billy's rock does not touch the window. Without fleshing out the notion of comparative naturalness at play, it is simply hard to assess what causes what.

9 Normality

All the counterfactual accounts of causation have so far in common that counterfactual dependence between actual events and absences is sufficient for causation. As a consequence, they all count too many omissions as causes. If Putin had watered my plant, it would not have died. Indeed, if the Queen of England had watered my plant, it would not have died. The same is true of anyone who didn't water my plant. This is an unwelcome result which questions the sufficiency of counterfactual dependence for causation.

One might block the unwelcome result by restricting causation to difference-making between occurring events.[9] This solution requires an ontological distinction between events and absences. Cases of prevention are then non-causal. Preventing an accident, for example, would not be

[9]One should presumably not require that all elements in a chain of counterfactual difference-making must be occurring events; at least not without defusing Schaffer's (2000) argument that there are cases of genuine causation, where a cause is related to its effect via absences.

causing the accident to be absent. And omissions would be non-causes in general, as defended by Beebee (2004). But it seems that preventing an event from happening is nothing but causing it to be absent. And some omissions seem to be causes while others are not.[10]

Another way to avoid causal omissions is to say that only occurring events can be causes while both events and absences can be effects. Then preventers may be causes, but omissions are always non-causes. My neighbour's failure to water my plant would then not cause it to die—even though she promised to water it. We are convinced by the argument of McGrath (2005) which establishes that the causal status of omissions depends on normality considerations.

The updated and modified Halpern-Pearl definitions can and have been amended by a condition of normality (Halpern, 2016, pp. 79-81&90-1). A normality order over possible worlds allows him to represent the different views about the causal efficacy of omissions to be found in the literature. The amended accounts understand causation roughly as de facto—or de counterfacto—dependence *witnessed by a possible world which is at least as normal as the actual one*. If Putin had watered my plant, it would not have died. True—but the world witnessing the counterfactual is less normal than the actual world, where Putin doesn't water my plant. By contrast, the world in which my neighbour waters the plant is at least as normal as the actual world, where she does not. Hence, my neighbour's failure to water the plant is a cause of its dying, whereas Putin's failure is not. The omission scenario poses no longer a problem due to the normality condition.

However, Putin's omission shows that counterfactual dependence between actual events and absences is no longer sufficient for causation on the amended definitions. The amendment by the normality condition breaks with the widely shared tradition in the wake of Lewis's (1973a) work that non-backtracking counterfactual dependence is sufficient for causation.

The amended definitions can also solve bogus prevention. This scenario

[10]The view that preventers and omissions are no genuine causes but may still figure in true counterfactual claims about genuine causation has been defended by Dowe (2001).

is isomorphic to overdetermination. The updated definition is a simple causal model account, meaning that it only takes structural equations and variable values into account. Hence, it must wrongly identify bodyguard's administering of antidote and assassin's refraining from poisoning target's coffee as individual causes of target's survival. Crucial is the de counterfacto conditional "if Bodyguard had not administered the antidote and assassin had put in the poison, target would have died". The world witnessing the de counterfacto dependence is the one where Bodyguard does not put in the antidote, assassin puts in the poison and target dies. We have no clear intuition whether this world is more or less normal than the actual world. Halpern (2016, pp. 88n) uses this lack of clarity: under the assumption that not putting anything in target's coffee is most normal, he declares the actual world *incomparable* to the witness world, and so bodyguard's administration of antidote is no cause of target's survival on the amended updated definition.

The result is similar for the amended modified definition. The above de counterfacto conditional expresses at the same time the simple counterfactual dependence of target's survival on the set containing both bodyguard's and assassin's actions. The witness world is the same and it was assumed to be incomparable to the actual world. Hence, bodyguard's administering of antidote is not part of a cause of target's survival on the amended modified definition.

One must wonder, however, why the witness world is not just as normal as the actual world and so at least as normal as the actual world. This would also explain the lack of clarity whether the witness world is more or less normal than the actual world. Besides this worry, the strategy to compare the normality of worlds looks promising so far.

On both Halpern-Pearl definitions, the dislodged boulder counts as a cause of the hiker's remaining unscathed. The reason is the de facto counterfactual "if the boulder had not been dislodged but it would still roll toward the hiker, the hiker would have been hit". It seems that the witness world—where the boulder has not been dislodged, but still rolls toward the hiker, the hiker does not duck, and so the boulder hits her—is less normal than the actual world. After all, it is a causally impossible world. Halpern (2016, p. 80) concurs by assuming that worlds which satisfy the

structural equations are more normal than worlds which do not. If so, the dislodged boulder does not count as a cause of the hiker's remaining unscathed. Analogously, the amended definitions say, under the assumption that causally impossible worlds are less normal than the actual, that flipping the switch in the respective switching scenarios is not a cause of the train's arrival. This seems like a win for the amended definitions!

Let's not be too hasty, however. Consider a modification of the causal model of the boulder scenario by replacing the structural equation of B with $B = F \vee F'$ and adding $\neg F'$ to the set V of literals—or set of value assignments if you prefer. $\neg F'$ stands for the absence of another boulder being dislodged. This modification just adds an absence to the causal model and so does not change the actual scenario. Hence, the dislodged boulder (F) is still not a cause of the hiker's remaining unscathed ($\neg E$)—as our analysis says. And yet, F is now a cause of $\neg E$ on the amended updated definition. Crucial is the de counterfacto conditional "if the boulder had not been dislodged but another boulder had been dislodged, the hiker would have been hit": $\neg F \wedge F' \mathrel{\Box\!\!\rightarrow} E$. In the witness world, the boulder is not dislodged, the other boulder is dislodged, a boulder rolls toward the hiker, the hiker does not duck, and so is hit by a boulder. This witness world is causally possible and seems at least as normal as the actual world. It follows from the same conditional and witness world that the dislodged boulder F is part of a cause of the hiker's remaining unscathed $\neg E$ on the amended modified definition—the other part is the absence of another boulder $\neg F'$. This seems odd.

The modified definition amended by a normality condition also faces troubles in the preemption scenarios. Recall that de facto accounts identify Suzy's throw as a cause of the window's shattering because her throw makes a counterfactual difference to the window's shattering when holding fixed that Billy's throw does not touch the window. In the witness world, Suzy does not throw, Billy does with unfailing accuracy, but his rock somehow does not touch the window, neither does Suzy's, and so the window does not shatter. It seems that this witness world is less normal than the actual world because it violates the structural equations. A natural solution to preemption is lost.

Suzy's throw still counts as part of a cause of the window's shattering—the

other part being Billy's throw. The counterfactual "if neither Suzy nor Billy had thrown a rock, the window would not have shattered" is true. Nothing happens in the causally possible witness world. This world seems to be at least as normal as the actual world. Hence, Billy's throw is part of a cause of the window's shattering on the amended modified definition. The updated definition amended by the normality condition has no such devastating consequence. The latest development for counterfactual accounts is to move from a normality order to the transmission of deviancy. We turn to such an approach next.

10 Counterfactual Transmission of Deviancy

To date, Gallow (2021) offers the perhaps most elaborate causal model theory of causation. His basic idea is that a cause transmits deviancy to its effect. He outlines his theory as follows:

> the theory says that C caused E whenever both C and E are deviant or non-inertial events, and there is an uninterrupted process which *transmits* C's deviancy to E. (p. 47)

The theory centres on a notion of causal network which is meant to represent the uninterrupted process transmitting deviancy from cause to effect. A process is, roughly speaking, uninterrupted if each step of the process depends only on its previous step. A network is only causal if each of its variables counterfactually depends on its direct predecessors if there are any. More formally, Gallow (2021, p. 83) proposes that

> each member of a set X of value assignments to the variables in $\mathcal{X}$ is a cause of a value assignment E to the variable E in a causal model $\langle M, V \rangle$ iff there is a minimal causal network in $\langle M, V \rangle$ which leads from the set $\mathcal{X}$ of variables to the variable E and assigns contrasts to the variables in $\mathcal{X}$ and E which are more default than their values.

Let us explain. A *network* $\mathcal{N}$ from a set $\mathcal{X}$ of variables to the variable E contains at least one directed path leading from C to E for each $C \in \mathcal{X}$. We may think of a network as a union of directed paths each of which leads from some $C \in \mathcal{X}$ to E. A network $\mathcal{N}$ is in a causal model $\langle M, V \rangle$ iff all of the variables of the network are in V.

A network $\mathcal{N}$ from a set $\mathcal{X}$ of variables to the variable E is *causal* iff there is an assignment K of contrast values to the variables in $\mathcal{N}$ such that all of the following conditions are true:

(a) E's contrast (value) is distinct from its (actual) value.

(b) The value of each variable D in the network $\mathcal{N}$ but not in $\mathcal{X}$, rather than its contrast, locally depends on D's $\mathcal{N}$-parents's values, rather than their contrasts.

(c) Every departure and return variable in $\mathcal{N}$ has a value which is more deviant than its contrast (Gallow, 2021, p. 77).

(a) constrains the assignment K of contrast values so that the contrast of the effect variable E must differ from its actual value. However, there is no constraint on the other variables in the network: variables other than E may have "contrast" values even though these "contrasts" are identical to their actual values.

(b) requires to explain some terminology. A variable A is one of D's $\mathcal{N}$-parents iff there is a directed edge $A \rightarrow D$ in $\mathcal{N}$. We may think of a variable's $\mathcal{N}$-parents as its parents within the network $\mathcal{N}$, as opposed to parents in the causal model.

Local dependence is defined via local models. We may think of a local model at E as the restriction of a causal model to the structural equation of E. Given a causal model $\langle M, V \rangle$, a *local model* $\langle M, V \rangle | E$ at E is $\langle E = \phi, V_E \rangle$, where $V_E \subseteq V$ assigns values only to the variable E and its parents. Note that E and its parents take on the same values as in the original causal model. Local dependence is then defined as follows:

> E, rather than $\neg E$, locally counterfactually depends on C, rather than $\neg C$, in $\langle M, V \rangle$ iff $\langle M, V \rangle | E \models C \wedge E$ and $\langle M, V \rangle | E[\neg C] \models \neg E$ (Gallow, 2021, pp. 69n).

The *rather-than* clauses are not necessary for local dependence when we only look at binary variables and assume that the contrast values are different from the actual values. These clauses *do* become important when checking whether there is an assignment K of "contrast" values such that condition (b) is satisfied. For then, there may be such assignments where the "contrast" values are identical to the actual values. This is how Gallow formalizes the uninterrupted process.

(c) requires that each departure and return variable in the network $\mathcal{N}$ has a value which is more deviant than its contrast. A variable A is a departure variable in $\mathcal{N}$ and B one of its return variables if there is a directed path $A \rightarrow ... \rightarrow B$ whose directed edges are not in $\mathcal{N}$. Gallow's account of deviancy says that firing neurons are more deviant than non-firing ones and this is close to all it says.

Finally, a causal network $\mathcal{N}$ from a set $\mathcal{X}$ of variables to a variable E is *minimal* iff there is no proper subnetwork $\mathcal{M}$ of $\mathcal{N}$ which is causal. A network $\mathcal{M}$ from a set $\mathcal{X}'$ of variables to the variable E is a proper subnetwork of $\mathcal{N}$ from $\mathcal{X}$ to E iff $\mathcal{X}' \subseteq \mathcal{X}$ and $\mathcal{M} \subset \mathcal{N}$. $\mathcal{M} \subset \mathcal{N}$ means that $\mathcal{N}$ contains all directed edges of $\mathcal{M}$ and at least one which is not in $\mathcal{M}$.

10.1 A Cause or Joint Causes?

Let us illustrate Gallow's theory of causation by applying it to the neuron scenario of overdetermination. The individual firing of each of the neurons C and A is sufficient for neuron E to fire. Following Mackie (1965) and Lewis (1986), Gallow thinks that "intuition is split" on whether C's firing is a cause of E's. Indeed, he is inclined to negate this and writes: "My theory will not say that C's firing individually caused E to fire." (p. 66) He points out that each overdeterminer alone does not *all by itself* cause the effect in "another case with a similar structure". (p. 64) One must wonder, however, whether the other case has a similar enough structure to the neuron scenario of overdetermination. The individual overdeterminers in the other case are not sufficient for the effect, whereas each individual overdeterminer is supposed to be sufficient for the effect in the neuron diagram of overdetermination. The latter just seems to mean that each overdeter-

miner *all by itself* is a cause of said effect.

The only network from the singleton $\{C\}$ to the variable E in overdetermination is $C \rightarrow E$. This network is not causal: there is no assignment of contrast values to C and E such that E locally depends on C and E's contrast value is distinct from its actual value. The reason is, of course, that A is still firing which is all by itself sufficient for E's firing. We have thus seemingly shown that the individual overdeterminer C is not a cause of E on Gallow's theory. Due to the symmetry of the scenario, the same reasoning applies to neuron A.

Gallow (2021, p. 66-8) thinks this is no problem as long as the joint firing of C and A is a—if not *the*—cause of E's firing. And indeed, there is a causal network from $\{C, A\}$ to the variable E: $C \rightarrow E \leftarrow A$. For this to be seen, assign C, A, and E the contrast values $\neg C$, $\neg A$, and $\neg E$. Then E's contrast differs from its actual value and its actual value, rather than its contrast, locally depends on the values C and A of its $\mathcal{N}$-parents's, rather than their contrasts. Moreover, all the assigned contrast values are more default than the variables's actual values. Finally, the causal network is minimal because the proper subnetworks are not causal, as we have seen in the paragraph above. Hence, C and A *jointly cause* E—or so says Gallow.

The problem of overdetermination motivates that candidate causes are members of a set X of value assignments: causes are joint causes—except in the case where the set X is a singleton. In tension with this motivation, however, Gallow's theory does not say that C and A are *joint* causes of E. The theory merely says that each member of $\{C, A\}$ is a cause of E—no qualification added. His own theory does not make explicit his own inclination that individual overdeterminers are no full causes. And there is a good reason for not doing so. In the neuron scenario of conjunctive causes, neuron C's firing does not *all by itself* cause neuron E to fire. For the neuron A must fire alongside C for E's firing. And yet, Gallow's theory deems here the singleton $\{C\}$ a cause of E, as well as the singleton $\{A\}$. Only the joint firing $\{C, A\}$ alone is sufficient for the effect and so *all by itself* a cause of it. However, $\{C, A\}$ does *not* count as a joint cause of E by minimality.

We have observed a tension between the motivation for considering sets of candidate causes and Gallow's theory. We think there is something off

with the motivation that C and A are only joint causes of E in the neuron diagram of overdetermination. Still, Gallow's theory obtains the desired verdicts which, however, go against his inclination: the individual overdeterminer C in the neuron diagram *does* in the end count as a cause simpliciter of E. The lesson we draw from this observation is that the most advanced counterfactual accounts using causal models *need* to consider a set, or an n-tuple, of causes to deal with overdetermination. While the motivation for this remains unclear to us, we think Halpern and Pearl (2005), Gallow (2021), and others may simply stipulate that candidate causes are members of a set of value assignments. It may be better to have no motivation rather than a questionable one.

10.2 Switches Revisited

Gallow's theory can handle an impressive set of scenarios including some switches. However, it has troubles with the *simple switch* (Andreas and Günther, 2024). Assume the action F of flipping the switch in the railroads is more deviant than not doing so. Then there is a minimal causal network leading from flipping the switch $\{F\}$ to the train's travelling on the right tracks R to the train's arrival E, namely the causal path $F \rightarrow R \rightarrow E$. Within this causal network, E locally depends on R, and R locally depends on F. E's contrast value $\neg E$ differs from it actual value. And the only departure variable F—departing from the minimal network to the variable L representing whether the train travels on the left tracks—and the only return variable E are both more deviant than their contrasts $\neg F$ and $\neg E$. Hence, flipping the switch counts as a cause of the train's arrival on Gallow's account if the flipping is more deviant than not. The problem applies also to the basic, the realistic, and the realistic basic switch (see Andreas and Günther (forthcoming, Chs. 4&6)).

Gallow (2021, p. 83) remarks that switches do not transmit deviancy to any effect. Indeed, if the action F of flipping the switch, or alternatively the position of the switch, is not more deviant than its respective contrast, then F is not a cause of the train's arrival E. However, F is then also not a cause of the train's travelling on the right tracks R. For the minimal causal network $F \rightarrow R$ does then not assign a contrast to F which is more default than its

actual value. It seems that Gallow's theory must take one hit here: either the deviant flipping of the switch causes the train's arrival, or else the default flipping—alternatively the switch's non-deviant position—does not cause the train to travel on the right tracks.

In general, only deviant events can be causes on Gallow's theory. If a position of an arbitrary switch is just as deviant as its alternative position, its actual position cannot be a cause at all. It is a hard pill to swallow that a *just as deviant* position of the switch cannot cause the train to go on the right tracks—and a harder pill that there are no cases where just as deviant positions of a switch cause anything.

It seems that at least some switches are *asymmetric*, meaning that one position is more deviant than the other. Just like a deviant action of flipping the switch in *simple switch*, such asymmetric switches transmit deviancy to a final effect on Gallow's theory. Here is an example. Suppose Paula wants to cross a mountain chain by car. There are two options. First, she can take a tunnel. Second, she can take an older road, leading via a mountain pass. By default, Paula takes the tunnel because it is the faster option.

The status of the tunnel—whether it is open or closed—acts like an asymmetric switch. If the tunnel is in the "default position" of being open, Paula takes the tunnel road. If the tunnel is in the "deviant position" of being closed, she takes the mountain pass. Either way, she crosses the mountain chain. Clearly, a closure of the tunnel causes Paula to go via the mountain pass. But we do not want to say that such a closure is a cause of her crossing the mountain chain. Our analysis agrees with both causal judgments and so solves asymmetric switches as well. Gallow's theory, by contrast, must say that the closure of the tunnel is a cause of Paula's crossing the mountain chain.

10.3 Gapless Transmission of Deviancy and Preventions

Gallow's main motivation for his theory is that there is an uninterrupted process which transmits deviancy from cause to effect. Such a process is naturally understood as a process on which deviancy is transmitted on each step from one variable value to the next. Gallow's causal networks,

however, do not demand such a gapless transmission of deviancy. Indeed, only the departure and return variables are required to have more deviant values than their contrasts.

Gallow (2021, p. 80) even shows how the transmission of deviancy without any gaps would look like in his counterfactual theory. He does so by the notion of a *productive network*. A productive network is a causal network in which every variable must have a value which is more deviant than its contrast. In a productive network, deviancy is transmitted on each step: each deviant value locally depends on the deviancy of its parents's values in the network.

Why does Gallow opt for his causal networks rather than the more natural understanding of transmission of deviancy in productive networks? One cost of the productive networks is that canonical double preventers are not causes. In the canonical model of double prevention, neuron C prevents neuron D from firing, which otherwise would have prevented E from firing. The neuron diagram contains the subgraph $C \rightarrow D \rightarrow E$ which is a causal network. But the intermediate variable D takes on the default value $\neg D$. The double preventer C is thus no cause of E on the natural understanding because it does not transmit deviancy to the variable D in an uninterrupted process—at least on Gallow's account of deviancy. It seems as if Gallow is not willing to sacrifice extensional adequacy for staying true to the more natural understanding of deviancy transmission as gapless.

Gallow's theory transmits deviancy only in the following sense: causes and effects must take on values more deviant than their contrasts and so must the departure and return variables. Any causal model specifies, for each variable, which values are more default than which others. Such a specification is invariant across all possible value assignments: if a value of a variable is more default than another, then it remains so in all counterfactual possibilities. The requirement that causes must be deviant leads to troubles with switching scenarios, as we have already seen. Another troublesome consequence of this requirement is that there is no causation by simple prevention. In a simple prevention scenario, some neuron C fires and thereby prevents another neuron E from firing. The effect neuron E does not fire and so takes on a default value on Gallow's account. This, on

its own, is sufficient to say on his theory that $\neg E$ has not been caused—a verdict which does not seem to accord with our commonsensical causal judgements.

Here is a proposal to remedy Gallow's situation using our agnostic states. For the simple prevention scenario, there is a state agnostic on whether the preventing neuron C and the effect neuron E fires but the would-be producer D of E still fires. In this agnostic state, we belief D fires and do not expect C to fire so that we expect E to fire. So E is default in the uninformative causal model. Assuming C then transmits its deviancy to $\neg E$. This account of deviancy can solve prevention because such an account of deviancy is relative to the context of an agnostic state. Gallow's account of deviancy, by contrast, has no context-relativity: a non-firing neuron is and remains default.

The different account of deviancy could also capture the double preventer C as a cause of E. On this account of deviancy, we expect D to fire in the relevant agnostic state. Assuming C then transmits its deviancy in this state to $\neg D$, which in turn transmits its deviancy to E. Gallow's non-relative account of deviancy cannot say that the non-firing of D is deviant. If his account could say so, he may have chosen productive networks over causal ones. It seems that a context-sensitive account of deviancy could help to account for double prevention while staying true to a notion of gapless transmission of deviancy.

Gallow (2021, p. 47) admits that his account of deviancy is incomplete. And he anticipates a more nuanced account of deviancy in footnotes 12 and 53. We hope that our outlined account of deviancy may provide some inspiration for completing the account of deviancy which would make Gallow's theory even stronger. At the very least, such an account would solve the problems with prevention and double prevention his theory faces.

10.4 Variants of Gallow's Theory

Gallow (2021, p. 87) proposes a variant of his theory of causation, namely his theory without the deviancy requirement imposed on the cause and ef-

fect variables. He does so in particular because of the problem that there is no causation by simple prevention and more generally in response to the problem that default events "can be neither causes nor effects". Causation then merely requires that there is a minimal causal network which leads from the set of cause variables to the effect variable. This modified theory does not stay true to the motivation of causation as transmission of deviancy. Furthermore, it says that *all* omissions, be they deviant or default, count as causes. Another difficult verdict ensues: just as deviant positions of switches count now as causes of final events like the train's arrival.

As Gallow (2021) mentions, there are two further options to modify his theory. First, causes must be deviant, but not effects. On this variant, simple preventers come out as causes and so do deviant omissions. We have therefore suggested to adopt this modification in Andreas and Günther (2024). Second, effects must be deviant, but not causes. On this variant, the problem of simple preventers persists. On both variants, Gallow's current account of deviancy cannot claim any transmission of deviancy.

We have seen that there are quite some tensions between the several theories of causation Gallow discusses and his main motivation that causes transmit their deviancy to their effects in an uninterrupted process. Given Gallow's current account of deviancy, none of his variants can say that causation is transmission of deviancy and that there is causation by prevention. Suppose there is causation by prevention. Then we have a "default effect" on his theory and so no transmission of deviancy to this "effect"—independently of whether transmission of deviancy is understood in the stepwise sense of productive networks or the gappy sense of causal networks. Conversely, if causation is transmission of deviancy, then there is no causation by prevention—at least on his current account of deviancy.

11 Trumping Preemption Revisited

In Andreas and Günther (forthcoming, Ch. 3, Sec. 7), we have shown that our analysis solves trumping preemption. We have done so with a simple causal model including only binary variables and no variables unmentioned in Schaffer's (2000) example. Thereby we have followed a method-

ological desideratum of Hitchcock (2001, p. 299) who wants to reproduce "our causal judgments without introducing events or variables beyond those explicitly presented in the various scenarios". Our solution is remarkable because we are not aware of any counterfactual account which can reproduce our causal judgments in trumping preemption with binary variables and "without introducing events or variables beyond those explicitly presented in" trumping preemption:

> the major and the sergeant stand before the corporal, both shout "Charge!" at the same time, and the corporal decides to charge. Orders from higher-ranking soldiers trump those of lower rank. (Schaffer, 2000, p. 175)

Without adding unmentioned variables or variable values, the structural equation of the trumping scenario is $E = C \vee (A \wedge \neg C)$: the corporal or soldier advances (E) just in case the major gives the command to advance (C), or the sergeant does (A) and the major does not ($\neg C$).

The extant causal model accounts rely on a purely semantic account of structural equations. Hence, they cannot distinguish the genuine cause C from the trumped cause A. The underlying reason is that, for them, the structural equation of the major-sergeant scenario is indistinguishable from the structural equation $E = C \vee A$ known from the overdetermination scenario (Hitchcock, 2001; Halpern and Pearl, 2005; Halpern, 2015; Gallow, 2021). Perhaps this situation can be remedied by a hyperintensional semantics for structural equations. But, as it stands, the extant causal model accounts face a problem which derives from the very foundation of their framework—the semantics of their structural equations.

The other counterfactual accounts fare no better. If the major had not shouted "Charge!", the corporal would still have charged. Hence, the simple counterfactual account says that the major's command is not a cause of the corporal's charge. Lewis's (1973a) analysis inherits this incorrect verdict, and so does the simple de facto account and Yablo's (2002). Only the simple de counterfacto account correctly says that the major's command is a cause of the corporal's charge. Hold fixed the non-actual absence of the sergeant's command. Then the corporal's charge counterfactually depends on the major's command. However, the simple de

counterfacto account misclassifies the sergeant's command as a cause of the corporal's charge. Hold fixed the non-actual absence of the major's command. Then the corporal's charge counterfactually depends on the sergeant's command.

Indeed, the only other account of causation we are aware of and which delivers the correct verdicts is Mackie's (1965) INUS account. The reason is that the major's command (C) is an instantiated INUS condition for the corporal's charge (E), while the sergeant's command (A) is not co-instantiated with the absence of the major's command ($\neg C$). Hence, $A \wedge \neg C$ is *not* an instantiated INUS condition for E. Mackie's INUS account is thus able to discern the genuine cause from the trumped cause, and trumping preemption without additional variables or variable values from overdetermination.

In response to the trouble with the minimalist causal model of trumping preemption, several authors have proposed different causal models for Schaffer's scenario. Halpern and Pearl (2005, p. 874), for example, assume "for definiteness that the sergeant and the major can each either order an advance, order a retreat, or do nothing." But they only obtain the desired results if they add a variable representing whether or not the sergeant's order is effective (p. 875). They add to the causal model an unmentioned absence—the sergeant's order not being effective. Indeed, they admit being unable to "speak about trumping preemption in [their] framework without being explicit as to how the trumping takes place." (ibid.)

Gallow (2021) similarly assumes that the major and the sergeant can each either do nothing, order to advance, or to stay put. It is interesting to see that Gallow's major and sergeant cannot order to retreat but to stay put instead. One must wonder how Halpern and Pearl and Gallow determine what possibilities are included in the causal model. After all, the new possible orders are not mentioned in the original scenario. Moreover, Gallow's specific model does not provide unambiguous truth values to some counterfactuals. For example, if the major had not shouted "Advance!" but the Sergeant still had, it is on Gallow's model unclear whether the soldier still would have advanced. For it might have been that the major then commands to stay put. This is a clear deviation from Schaffer's original scenario.

Our minimalist model of trumping preemption does not assume anything that remains unmentioned in the original scenario. Unlike Halpern and Pearl, no unmentioned variable is added. And unlike Gallow, we need not assume the additional possibilities that the major either does nothing or shouts "Stay put!" Like Mackie's INUS account, our analysis may thus serve as a proof of concept that such additions are not required to solve trumping preemption.

12 Conclusion

We have looked at accounts motivated by the idea that causes are difference-makers. The simple counterfactual account succumbs to the problem of redundant causation. This motivates dropping the necessity of counterfactual dependence between actual events and absences for causation. The simple de facto and de counterfacto accounts do so but retain its sufficiency for causation. As a consequence, all omissions count as causes. This motivates also dropping the sufficiency of counterfactual dependence between actual events and absences. The amended causal model accounts, which rely on comparing the normality between possible worlds, drop both necessity and sufficiency of counterfactual dependence. One must wonder what remains of the guiding idea that causes *are* counterfactual difference-makers.[11]

We have seen that refined accounts in terms of counterfactual difference-making solve the problem posed by redundant causation. However, the extant solutions create new problems. Lewis's (1973a) imposition of transitivity on counterfactual difference-making to solve early preemption creates the problem that the dislodged boulder counts as a cause of the hiker's remaining unscathed. The de facto and de counterfacto accounts of Yablo (2002), Hitchcock (2001), Halpern and Pearl (2005) and Halpern (2015) provide the same problematic verdict among others. Amending the Halpern-Pearl accounts by a condition of normality solves the boulder scenario but not a factual equivalent thereof. The extant solutions to the problem of

[11] If one wants to retain the idea of causes as difference-makers, one should perhaps look into the notion of factual difference-making offered by Andreas and Günther (2025).

redundant causation consistently lead to new problems. In this sense, redundant causation still haunts accounts in terms of counterfactual dependence.

Switching scenarios pose another problem for most counterfactual accounts. Of the accounts we considered only the Halpern-Pearl definitions amended by a certain normality condition obtain the desired results in the simple, the basic, the realistic, and the realistic basic switch. However, the amended definition of Halpern (2015) then fails for early preemption: Billy's throw is counted as part of a cause of the window's shattering—the other part being Suzy's throw. Notably, the amended definition of Halpern and Pearl (2005) can still solve early preemption.

We are not aware of *any* counterfactual account that can reproduce our causal judgments in our minimalist model of the trumping scenario. The causal model accounts due to Hitchcock (2001), Halpern and Pearl (2005), Halpern (2015), and Gallow (2021), in particular, rely on a purely semantic account of structural equations. Hence, they cannot distinguish the genuine cause from the trumped cause in our minimalist model. The underlying reason is that, for them, the structural equation of trumping is indistinguishable from the structural equation of the overdetermination scenario. All the extant causal model accounts using Pearl's (2009) framework of causal models face a problem here which derives from the very foundation of their framework—the semantics of their structural equations.

We have refrained from discussing entanglements in this lengthy comparison chapter. One reason is that Beckers (2021) has already shown that the Halpern-Pearl definitions deliver unintuitive results for his series of six scenarios of entangled causes. The criticism applies as well to Hitchcock's account. Lewis's (1973a) analysis wrongly and unsurprisingly says that the conjunctive factor of an overdeterminer is never a genuine cause. Yablo's (2002) account is at best unclear. And while Gallow's (2021) theory accounts well for the first five scenarios, it fails for the sixth, where the conjunctive factor is not a genuine cause. His theory says that the conjunctive factor is a genuine cause—a joint cause with the disjunctive cause. We have shown in Andreas and Günther (forthcoming, Ch. 5) that our analysis delivers the intuitive verdicts for all scenarios of entanglement studied by Beckers (2021).

Gallow's (2021) theory faces two main problems. First, it has troubles with the switching scenarios mentioned above. Second, his theory can either account for causation by prevention, or else stay true to its motivation that causation is counterfactual transmission of deviancy—but not both. It seems to us that the two problems may be overcome by a different account of deviancy not unlike the one we outlined.

Our analysis of causation, by contrast, accounts for all the causal scenarios mentioned. At least with respect to this set of scenarios, our analysis tallies best with our causal judgments. Our analysis, furthermore, does not suffer from internal conceptual tensions and stays true to its guiding idea: there must be an active path leading from a cause to its effect in a causal model agnostic on both.

Finally, we will show that our causal model analysis is well-founded by a reductive theory of causation. This foundation will answer, at least in part, to the question what causal models are appropriate for determining causation. We turn to the reductive theory in Part II of Andreas and Günther (forthcoming), and explain how it grounds the causal model analysis in the Conclusion and Synthesis of this book.

References

Andreas, Holger and Günther, Mario (2021). Difference-Making Causation. *Journal of Philosophy* **118**(12): 680–701.

Andreas, Holger and Günther, Mario (2021). A Ramsey Test Analysis of Causation for Causal Models. *The British Journal for the Philosophy of Science* **72**(2): 587–615.

Andreas, Holger and Günther, Mario (2024). A Lewisian Regularity Theory. *Philosophical Studies* **181**(9): 2145–2176.

Andreas, Holger and Günther, Mario (forthcoming). Actual Causation. *dialectica* .

Andreas, Holger and Günther, Mario (2025). Factual Difference-Making. *Australasian Philosophical Review* **x**(x): x–y.

Andreas, Holger and Günther, Mario (forthcoming). *From Reasons to Causes: A Theory of Causation*. Cambridge University Press.

Beckers, Sander (2021). Causal Sufficiency and Actual Causation. *Journal of Philosophical Logic* **50**(6): 1341–1374.

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

Coady, David (2004). Preempting Preemption. In *Causation and Counterfactuals*, edited by J. Collins, N. Hall, and L. Paul, MIT Press. 325–340.

Dowe, P. (2001). A Counterfactual Theory of Prevention and 'Causation' by Omission. *Australasian Journal of Philosophy* **79**(2): 216–226.

Gallow, Dmitri J. (2021). A Model-Invariant Theory of Causation. *The Philosophical Review* **130**(1): 45–96.

Hall, Ned (2004). Two Concepts of Causation. In *Causation and Counterfactuals*, edited by J. Collins, N. Hall, and L. Paul, MIT Press. pp. 225–276.

Hall, Ned (2007). Structural Equations and Causation. *Philosophical Studies* **132**(1): 109–136.

Halpern, Joseph (2016). *Actual Causality*. Cambridge, MA: MIT Press.

Halpern, Joseph Y. (2015). A Modification of the Halpern-Pearl Definition of Causality. *Proc. 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)* : 3022–3033.

Halpern, Joseph Y. and Hitchcock, Christopher (2010). Actual Causation and the Art of Modeling. In *Heuristics, Probability, and Causality: a Tribute to Judea Pearl*, edited by R. Dechter, H. Geffner, and J. Y. Halpern, London: College Publications. 383–406.

Halpern, Joseph Y. and Pearl, Judea (2001). Causes and Explanations: A Structural-Model Approach, Part I: Causes. In *In Proc. Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI 2001)*. 194–202.

Halpern, Joseph Y. and Pearl, Judea (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal for the Philosophy of Science* **56**(4): 843–887.

Hitchcock, Christopher (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *The Journal of Philosophy* **98**(6): 273–299.

Lewis, David (1973a). Causation. *The Journal of Philosophy* **70**(17): 556–567.

Lewis, D. (1973b). *Counterfactuals*. Oxford: Blackwell.

Lewis, David (1986). Postscripts to "Causation". In *Philosophical Papers. Volume II*, edited by D. Lewis, Oxford University Press. pp. 172–213.

Lewis, David (2000). Causation as Influence. *The Journal of Philosophy* **97**(4): 182–197.

Mackie, J. L. (1965). Causes and Conditions. *American Philosophical Quarterly* **2**(4): 245–264.

McGrath, Sarah (2005). Causation By Omission: A Dilemma. *Philosophical Studies* **123**(1-2): 125–148.

Paul, Laurie and Hall, Ned (2013). *Causation: A User's Guide*. Oxford.

Pearl, Judea (2000). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press, 1st edn.

Pearl, Judea (2009). *Causality: Models, Reasoning and Inference*. New York, NY, USA: Cambridge University Press, 2nd edn.

Ramachandran, Murali (1997). A Counterfactual Analysis of Causation. *Mind* **106**(422): 263–277.

Schaffer, Jonathan (2000). Trumping Preemption. *Journal of Philosophy* **97**(4): 165–181.

Woodward, James (2003). *Making Things Happen : A Theory of Causal Explanation*. Oxford University Press.

Yablo, Stephen (2002). De Facto Dependence. *The Journal of Philosophy* **99**(3): 130–148.

Yablo, Stephen (2004). Advertisement for a Sketch of an Outline of a Proto-Theory of Causation. In *Causation and Counterfactuals*, edited by N. Hall, L. A. Paul, and J. Collins, Cambridge: MIT Press. 119–137.