DOI: 10.1111/phpr.12925

Philosophy and Phenomenological Research

ORIGINAL ARTICLE

Causal counterfactuals without miracles or backtracking

J. Dmitri Gallow

Dianoia Institute of Philosophy, ACU

Correspondence
Dianoia Institute of Philosophy, ACU.
Email: dmitri.gallow@acu.edu.au

Abstract

If the laws are deterministic, then standard theories of counterfactuals are forced to reject at least one of the following conditionals: 1) had you chosen differently, there would not have been a violation of the laws of nature; and 2) had you chosen differently, the initial conditions of the universe would not have been different. On the relevant readings—where we hold fixed factors causally independent of your choice—both of these conditionals appear true. And rejecting either one leads to trouble for philosophical theories which rely upon counterfactual conditionals—like, for instance, causal decision theory. Here, I outline a semantics for counterfactual conditionals which allows us to accept both (1) and (2). And I discuss how this semantics deals with objections to causal decision theory from Arif Ahmed.

If the laws of nature are deterministic, then standard theories of counterfactuals are forced to deny one of the following:

(A1) Had you chosen differently, no law of nature would have been violated.
(A2) Had you chosen differently, the initial conditions of the universe would not have been changed.

On the relevant readings, where we hold fixed factors causally independent of your choice, both of these conditionals appear true. And denying either leads to trouble for philosophical

theories which rely upon counterfactual conditionals—like, for instance, causal decision theory.

In §1 below, I will explain more carefully why standard theories of counterfactuals are forced to deny one of these conditionals at deterministic worlds, and why this leads to problems for causal decision theory. Then, in §§2–3, I will outline a different semantics for counterfactual conditionals. And in §4, I will demonstrate that this semantics allows us to accept both (A1) and (A2). §5 concludes with some further discussion of the theory and what it has to say about conditionals like "Had you chosen differently, it would have been the case that *either* a law of nature was violated *or* the initial conditions were different".

1 | A PUZZLE ABOUT CAUSAL COUNTERFACTUALS

1.1 | Causal counterfactuals

A counterfactual conditional is a claim expressible by a sentence in the following form: if it were the case that P, then it would be the case that Q. I'll abbreviate conditionals like this with: $P > Q$. When the antecedent is false, evaluating a counterfactual conditional requires us to suppose that some part of the world is different, and then to work out what *else* about the world would have to change as a result.

In general, there will be many ways of doing this, depending upon which parts of the world we hold fixed and which we allow to vary. Suppose you stand on the top of a building with no safety net below. In this context, consider the following dialogue:

Me: We're high enough that, if you were to jump, you would die.
You: I don't have a death wish! If I were to jump, there would be a safety net, and I wouldn't die.

Neither of us appear to have said anything false, but I uttered a counterfactual of the form $J > D$, and you uttered a counterfactual of the form $J > \neg D$. Assuming a principle of conditional non-contradiction, these cannot both be true at once.[1] (The principle I have in mind is this: so long as P is possible, $\neg[(P > Q) \wedge (P > \neg Q)]$. I'll take this principle for granted throughout.)

The standard resolution is to acknowledge context-sensitivity in counterfactuals.[2] My claim held fixed the lack of a safety net, whereas yours held fixed your lack of a death wish. Because we held different things fixed, we made different claims. You said something about the necessary causal precursors of your jumping, holding fixed your lack of a death wish. Whereas I said something about the inevitable causal consequences of your jumping, holding fixed the lack of a safety net.

My focus here is on this second kind of counterfactual, which I will call a *causal* counterfactual. In general, causal counterfactuals hold fixed factors which are not causally influenced by the antecedent.[3] (Throughout, then, 'counterfactual' always means 'causal counterfactual', and '>' always stands for the causal counterfactual conditional.)

[1] This example comes from Jackson (1977, p. 9). Jackson's reaction to the case differs from my own; he suggests that you have confused an indicative conditional for a counterfactual—and that, interpreted as a counterfactual, your claim is false.

[2] Cf. Lewis (1979).

[3] Theories of causal counterfactuals like these have been explored by Jackson (1977), Galles & Pearl (1998), Woodward (2003), Kment (2006), Briggs (2012), Huber (2013), and Hiddleston (2005), among others.

1.2 | The puzzle

There is a puzzle about how causal counterfactuals interact with determinism. This puzzle puts pressure on us to deny some intuitive counterfactual judgements. And denying these intuitive judgements leads to powerful objections to philosophical theories formulated in terms of causal counterfactuals.

To appreciate this puzzle: say that laws are *deterministic* only if two metaphysically possible worlds satisfying those laws differ iff they have different *initial conditions*.[4,5] And, given some laws, say that there is a *miracle* iff one of those laws is violated.[6] Now, suppose the laws are deterministic and unviolated, so that, actually, there are no miracles. And suppose that you face a choice between two options, which I'll call '*a*' (for 'actual') and '*b*'. Actually, you choose *a*. Then, the puzzle is that each of the following claims are plausible, but they are jointly inconsistent.[7] (The fourth claim is schematic; to endorse it is to endorse all of the claims you get by substituting any claims for P, Q, R, and S.)

(B1) If you hadn't chosen *a*, there would not have been a miracle.
(B2) If you hadn't chosen *a*, the initial conditions would not have been changed.
(B3) For some true P, if you hadn't chosen *a*, P wouldn't have been true.
(B4) If $P > Q$, $P > R$, and $Q \wedge R$ metaphysically necessitates S, then $P > S$.

Pick any true proposition, P. Since the laws are deterministic, there is no metaphysically possible world at which the initial conditions are unchanged, there is no miracle, and P is false. So, the initial conditions being the same and there being no miracle metaphysically necessitates P. So (B1), (B2), and (B4) tell us that, if you hadn't chosen *a*, P would have been true. P was arbitrary, so the same goes for any true proposition P. So, for any true P, if you hadn't chosen *a*, P would have been true. And this contradicts (B3), given conditional non-contradiction.[8]

Standard semantics for counterfactual conditionals validate (B4). So they force a choice between denying (B1), denying (B2), and denying (B3). This is a difficult choice to make. If (B3) is false, then it is hard to understand the point of counterfactual thinking in a deterministic world. For, if (B3) is false and the world is deterministic, then nothing counterfactually depends upon your choice. Some incompatibilists may welcome this result, insisting that, in a deterministic world, counterfactual thinking has no role to play in rational deliberation. If you're determined to

[4] A possible world *satisfies* some laws iff the laws are true at that world. Likewise, laws are *violated* at a possible world iff the laws are false at that world.

[5] The *initial conditions* of the world are just some brief temporal interval at the beginning of the universe (or, in light of relativity: the past Cauchy development of a Cauchy surface near the beginning the universe).

[6] Lange (2000), Braddon-Mitchell (2001), and Kment (2006) have responded to variants of this puzzle by allowing that it is possible for the actual laws to remain laws at worlds where they are violated. (See Gibbs (2020) for criticism of this approach.) As I'm using the term 'miracle', whether a miracle occurs at another possible world does not depend upon whether the laws at the actual world are laws at that world. It only matters whether the actual laws are violated at that world or not.

[7] My presentation of the puzzle closely follows Dorr (2016).

[8] I am taking conditional non-contradiction for granted, but not everyone to discuss this puzzle has done so. In particular, in a letter to Jonathan Bennett, David Lewis proposed—but did not endorse—a semantics for counterfactuals which allows us to hold on to each of (B1), (B2), and (B3) by denying conditional non-contradiction along with the principle (B4). See Lewis (2020).

choose a, then there's no point to deliberating about whether to choose b instead. I disagree—but even if I concede that we are not free to do otherwise in a deterministic world, and that for this reason, counterfactuals have no role to play in rational deliberation in a deterministic world, we should still be able to adequately explain why things happen as they do. Whether these explanations are given freely is a separate question from whether the explanations are any good. But since many good scientific explanations appeal to counterfactual facts about how some things depend upon others,[9] denying (B3) would undermine our ability to adequately explain in a deterministic world.

At least when it comes to "standard" counterfactuals—which may or may not be *causal* counterfactuals, as I'm using the term here—Dorr (2016) denies (B2). Not only does he deny (B2), he *affirms* (C2):

(C2) If you hadn't chosen a, the initial conditions would have been changed.

To accept (C2) and similar counterfactuals is to say that, in deterministic worlds, causal counterfactuals regularly *backtrack*. As I'll use the term here, a causal counterfactual *backtracks* iff it says that, were things different at some time t, things would also have been different at some time earlier than t. For instance, on its true reading, your counterfactual "If I were to jump, there would be a safety net" backtracks. It says that, if you were to jump now, then there would have been a safety net *before* you jumped, even though, in fact, there is no safety net. (It's relatively uncontroversial that this counterfactual backtracks. What's more controversial is whether *causal* counterfactuals backtrack.) Dorr says that standard counterfactuals backtrack all the way to the initial conditions.[10]

Many think that (C2) follows from the negation of (B2). They accept the principle of conditional excluded middle, CEM, according to which there is no middle ground between $P > Q$ and $P > \neg Q$ (for any P, Q):[11]

$$\text{CEM} \qquad (P > Q) \vee (P > \neg Q)$$

However, not everyone accepts CEM. This opens up the possibility of denying (B2) without accepting (C2). Unfortunately, it doesn't appear to me that this offers a satisfying resolution of our puzzle. Those who think that there is a 'middle ground' between $P > Q$ and $P > \neg Q$ typically think that this middle ground is occupied by so-called "might" 'counterfactuals'—propositions expressed by sentences of the form 'if it were the case that P, it *might* have been the case that Q'.[12] And, on this view, the negation of (B2) commits us to (D2), which doesn't seem much better than (C2).

[9] See, e.g., Woodward (2003) for the role of counterfactual thinking in causal explanation.

[10] For Dorr's distinction between "standard" and "non-standard" counterfactuals, see p. 245. Similar positions are defended by Bennett (1984), Nute (1980, §5.3), Goggans (1992), Albert (2003), Kutach (2002), Loewer (2007), Wilson (2014), and Goodman (2015), among others.

[11] See, e.g., Stalnaker (1980) and DeRose (1999).

[12] Defenders of CEM say that these 'might' counterfactuals are just epistemic modals scoping over ordinary 'would' counterfactuals. That is, the logical form of 'if it were the case that P, it might have been the case that Q' is $\Diamond_e(P > Q)$, where '$\Diamond_e$' is an epistemic possibility operator.

(D2) If you hadn't chosen a, the initial conditions might have been changed.

If we accept (D2) and similar counterfactuals, we will say that 'might' counterfactuals regularly backtrack. But these backtracking 'might' counterfactuals seem false for the same reason that the backtracking 'would' counterfactuals seem false: it seems that we do not have *any* causal influence over the initial conditions. And so it seems that, had you not chosen a, the initial conditions *would not* have been any different; it's not the case that they *might* have been changed, had you chosen differently.

Lewis (1979) denies (B1).[13] Not only does he deny (B1), he *accepts* (C1):

(C1) If you hadn't chosen a, there would have been a miracle.

To accept (C1) is to say that your choosing differently would have been a miraculous, inexplicable event. It is to say that, had you chosen to not read this paper, that choice would have had profound implications for our best theories of fundamental physics. (Lewis *also* says that causal counterfactuals generally backtrack, though he thinks the backtracking is limited to the very recent past; Lewisian counterfactuals don't backtrack all the way to the initial conditions.)

Again, if we reject CEM, we could reject (B1) without accepting (C1). However, if we think that the 'middle ground' between $P > Q$ and $P > \neg Q$ is occupied by a 'might'-counterfactual, then we will still end up accepting (D1), which doesn't seem much better than (C1):

(D1) If you hadn't chosen a, there might have been a miracle.

A miraculous 'might' counterfactual like (D1) appears false for the same reasons that a miraculous 'would' counterfactual like (C1) appears false: whether you choose a or not doesn't appear to have any causal influence over whether the laws of nature are violated. Just as it seems incorrect to say that our best fundamental physical theories *would* have been false, had you chosen differently, it also seems incorrect to say that they *might* have been false, had you chosen differently.

So long as we adhere to the principle (B4), we are forced to deny one of (B1), (B2), and (B3). This is puzzling in part because each of (B1), (B2), and (B3) appears true when they are given a causal reading. So there's a puzzle for the semantics of English language counterfactuals. But there are additional puzzles for those of us who want to use causal counterfactuals in our philosophical theorising—whether or not we tether those counterfactuals to their English language counterparts. For instance, many of us appeal to something like causal counterfactual conditionals in our theorising about rational choice. We may not be bothered if these theoretical tools end up differing from the English conditionals which inspired them. But whatever we say about the connection between the English language and the causal counterfactuals used in philosophical theorising, so long as our theoretical tools satisfy (B4), we will have to deny one of (B1), (B2), and (B3) when the claims are understood in our favoured theoretical sense. And doing so will carry theoretical costs, quite apart from the counterintuitiveness of denying one the English sentences (B1), (B2), or (B3).

[13] Similar positions are defended by Jackson (1977), Halpin (1991), Lange (2000), Beebee (2003), and Kment (2006), among others.

1.3 | Causal decision theory

Take, for instance, causal decision theory (CDT). Many formulations of CDT utilise causal counterfactual conditionals.[14] Just for illustrative purposes, take the version of CDT from Gibbard & Harper (1978) and Stalnaker (1981). According to this theory, you should choose whichever act *would* bring about the best outcome, in expectation. That is, if $\mathcal{O}$ is the collection of potential *outcomes*, Pr your probability function, and $\mathcal{D}(o)$ the desirability of the outcome o, then you should choose whichever option x maximises $\mathcal{U}(x)$, where

$$\mathcal{U}(x) \stackrel{\text{def}}{=} \sum_{o \in \mathcal{O}} \Pr(x > o) \cdot \mathcal{D}(o)$$

In the discussion below, I'll appeal to a helpful fact about utility. To appreciate this fact, first define $\mathcal{U}(x \mid y)$ to be the utility that x has, conditional on you choosing y:

$$\mathcal{U}(x \mid y) \stackrel{\text{def}}{=} \sum_{o \in \mathcal{O}} \Pr(x > o \mid y) \cdot \mathcal{D}(o)$$

Here's the fact: in a choice between two options, a and b, if both $\mathcal{U}(b \mid a) > \mathcal{U}(a \mid a)$ and $\mathcal{U}(b \mid b) > \mathcal{U}(a \mid b)$—that is: if the utility of b exceeds the utility of a, conditional on both a and b—then the unconditional utility of b will exceed the unconditional utility of a, and so CDT will say that b is rational and a is irrational.[15]

Now, consider the following two decisions, adapted from Ahmed (2013, 2014a, 2014b):[16]

Betting on a miracle You are certain that the laws are deterministic, and that there are not and will never be any miracles. You are given a choice between two bets. Bet a pays out \$10 if there's no miracle and \$0 if there is. Bet b pays out \$1 if there's no miracle and \$11 if there is.

	There's no miracle	There's a miracle
Bet a	\$10	\$0
Bet b	\$1	\$11

Betting on the past You are certain that the laws are deterministic and that the initial conditions were c. You are given a choice between two bets. Bet a pays out \$10 if the initial conditions were c and \$0 otherwise. Bet b pays out \$1 if the initial conditions were c and \$11 otherwise.

[14] In addition to Gibbard & Harper (1978) and Stalnaker (1981), Lewis (1980) defines his causal dependency hypotheses in terms of counterfactual independence, and the imaging functions from Sobel (1994) and Joyce (1999) are explicated in terms of counterfactual conditionals.

[15] To see this, first note that $\mathcal{U}(x) = \sum_o \Pr(x > o) \cdot \mathcal{D}(o) = \sum_o \sum_y \mathcal{D}(o) \Pr(x > o \mid y) \cdot \Pr(y) = \sum_y \Pr(y) \cdot \sum_o \Pr(x > o \mid y) \cdot \mathcal{D}(o) = \sum_y \Pr(y) \cdot \mathcal{U}(x \mid y)$. Therefore, unconditional utility $\mathcal{U}(x)$ is a linear average of conditional utilities $\mathcal{U}(x \mid y)$, with weights given by your probability that you'll select the options y. So if the conditional utility for b is greater than the conditional utility for a, given every option, then the unconditional utility for b will exceed the unconditional utility for a, no matter what your option probabilities are.

[16] For additional discussion of decisions like these, see Solomon (2021) and Elga (2022).

	Initial conditions are c	Initial conditions are not c
Bet a	\$10	\$0
Bet b	\$1	\$11

It seems that, in both decisions, it is rational for you to take bet a and irrational for you to take bet b. But, if we deny (B1), (B2), or (B3), then CDT will disagree. (By the way, in the following, I will assume CEM, so that denying (B1) commits us to (C1), and denying (B2) commits us to (C2). I make this assumption in the interests of simplicity; similar troubles await even if we deny CEM.)

Suppose first that we deny (B1) and accept (C1). Then, CDT will give apparently bad advice in **Betting on a miracle**. There are two cases to consider. Either you take a or you take b. If you take a, then you're certain that, if you *were* to take a, there wouldn't be any miracle, and you'd get \$10.[17] And, if you *were* to take b, there would be a miracle, and you would win \$11. So $\Pr(a > \$10 \mid a) = \Pr(b > \$11 \mid a) = 100\%$. And so $\mathcal{U}(a \mid a) = \mathcal{D}(\$10)$ and $\mathcal{U}(b \mid a) = \mathcal{D}(\$11)$. Since \$11 is more desirable than \$10, $\mathcal{U}(b \mid a) > \mathcal{U}(a \mid a)$. On the other hand, suppose you take b. Then, were you to take b, there wouldn't be a miracle and you'd win \$1. And, were you to take a, there would be a miracle, and you'd win \$0. So $\Pr(b > \$1 \mid b) = \Pr(a > \$0 \mid b) = 100\%$. And so $\mathcal{U}(a \mid b) = \mathcal{D}(\$0)$ and $\mathcal{U}(b \mid b) = \mathcal{D}(\$1)$. Since \$1 is more desirable than \$0, $\mathcal{U}(b \mid b) > \mathcal{U}(a \mid b)$. So b has a higher utility than a, whether you take a or b. So CDT will say that b is rational and a is irrational. This looks like the wrong verdict.

Suppose on the other hand we deny (B2) and accept (C2). Then, CDT will give apparently bad advice in **Betting on the past**. Suppose you take a. Then, you'll expect to get \$10 from bet a, and you'll expect that, were you to take b, you'd get \$11 (since, were you to take b, the initial conditions would be different). So we'll again have that $\Pr(a > \$10 \mid a) = \Pr(b > \$11 \mid a) = 100\%$. So $\mathcal{U}(a \mid a) = \mathcal{D}(\$10)$ and $\mathcal{U}(b \mid a) = \mathcal{D}(\$11)$. Suppose on the other hand you take b. Then, you'll expect to get \$1 from b, and you'll expect that, were you to take a, you'd get \$0 (since, were you to take a, the initial conditions would be different). So again: $\Pr(a > \$0 \mid b) = \Pr(b > \$1 \mid b) = 100\%$, so $\mathcal{U}(a \mid b) = \mathcal{D}(\$0)$ and $\mathcal{U}(b \mid b) = \mathcal{D}(\$1)$. So b will have a higher utility than a, whether you take a or b. So CDT will say that b is rational and a is irrational.

Denying (B3) only makes matters worse. If we deny (B3), then *every* choice will always have exactly the same utility, and no option will ever be deemed irrational.

Some have responded to cases like these by proposing modifications to CDT.[18] Others have suggested that **Betting on a miracle** and **Betting on the past** are not genuine decisions,[19] or that the kinds of situations in which you could plausibly face these decisions are so outré that our judgements about rational choice in those decisions are not trustworthy.[20] From my perspective, it

[17] Here, I assume that, if you actually choose x, then were you to choose x, there wouldn't be a miracle. We could instead say that there are counterfactual miracles even when they aren't *needed* to make the antecedent true. But saying this only makes matters worse for CDT, in the sense that it will only *lower* the utility of bet a.

[18] See, for instance, Williamson & Sandgren (forthcoming), Sandgren & Williamson (2021), and Kment (ms).

[19] See Joyce (2016).

[20] See, for instance, Dorr (2016, §7)'s discussion of decisions like **Betting on the past**. It is also important to bear in mind the observation from footnote 37 of Dorr (2016): given some ways of presenting the proposition that the initial conditions are c (e.g., "the initial conditions are what they actually are"), there is no possibility in which bet a fails to pay out \$10, and no possibility in which bet b pays out more than \$1. If the bet is presented in these ways, then taking bet a will causally dominate taking bet b. I am assuming that it is possible for you to be very confident that the initial conditions are c even

would be better to reject the causal counterfactuals which lead CDT into trouble in these decisions. If we say that both miracles and the past are counterfactually independent of your choice, then CDT will advise you to take bet *a* in both decisions.

***

Each of (B1), (B2), and (B3) are very natural. Given that we are careful to understand them as *causal* counterfactuals, their negations appear false. If we must deny one of (B1), (B2), and (B3), then there are serious challenges to causal decision theory. Since I am inclined to accept a broadly causalist theory of rational choice, I would prefer a semantics for causal counterfactuals which denies (B4). I provide a semantics like this in Gallow (2016). In §§2 and 3 below, I will introduce and motivate this semantics. Then, in §4, I will show that the semantics allows us to accept (B1) and (B2)—it will obviously satisfy (B3). Finally, in §5, I will explain why the semantics violates the schematic principle (B4), and discuss what the theory has to say about counterfactuals like "If you hadn't chosen *a*, then it would have been the case that either there was a miracle or the initial conditions were different".

2 | CAUSAL INFLUENCE AND CAUSAL COUNTERFACTUALS

In my view, causal counterfactuals presuppose a system of *causal influence*. After all, what makes a counterfactual *causal* is that it holds fixed factors which are not causally influenced by the antecedent. It only allows to swing free those factors which are causally downstream of the antecedent. So, before we evaluate a causal counterfactual, we must understand how the antecedent fits into the world's causal structure: what influences it, and what it influences.

The reason I will accept the counterfactual (B1) is that I will deny that there is any causal influence running from whether you choose *a* to whether there is a miracle. Likewise, I will accept (B2) because I will deny that there is any causal influence running from whether you choose *a* to whether the initial conditions are different.

2.1 | Causal influence

Causal influence is a relation which holds between *variables*. Variables are the contrastive generalisation of events. For illustration, let us begin with the Lewisian view that events are properties of spacetime regions, or spacetime regions taken in intension. That is, a *Lewisian event*, *e*, is a class of possible spacetime regions. Spacetime regions belonging to the class are regions in which the event occurs; those not belonging to the class are regions in which it does not occur.[21] Corresponding to this class is a function from spacetime regions at possible worlds to $\{1, *\}$, where '$*$' is some arbitrary entity. If this function maps a region, *R*, to 1, then *R* is a region in which *e* occurs. If it maps *R* to $*$, then *R* is not a region in which *e* occurs. (The choice of both '1' and '$*$' is arbitrary. Any other choice would do just as well. What's important is how we divide up the possible spacetime regions—which we include and which we exclude—and not how we designate the included and the excluded.) Now, a *variable*, *V*, is a *contrastive* property of a spacetime region. Taking the

under a "non-cheesy" guise, one which would pick out a false proposition in nearby possible worlds where the initial conditions are different.

[21] See Lewis (1986a)

Lewisian view as our point of departure, we may say that a variable is a class of classes of possible spacetime regions. Spacetime regions belonging to one of the classes are regions in which the variable takes on a value; those not belonging to any of the classes are regions in which it does not take on a value. Spacetime regions which belong to the same class are alike with respect to the variable property V. Corresponding to this class of classes is a function from possible spacetime regions to $\mathbb{R} \cup \{*\}$. If this function maps a region, R, to a real number v, then the variable takes on a value in the region R, and that value is v. If the function maps R to $*$, then the variable does not take on a value in the region R. (Our choice of real numbers from $\mathbb{R}$ is arbitrary. What's important is how we divide up the possible spacetime regions, and not how we designate the cells of the division.)

Whereas events correspond to English expressions like "my throwing the ball", "the dinner", and "the game's end", variables correspond to expressions like "whether I throw the ball", "how much I eat at dinner", and "when the game ends". When variables causally influence each other, this is naturally expressed in English with the verbs "affects" and "influences". For instance: "whether I throw the ball affects when the game ends", and "how much I eat at dinner influences whether I throw the ball".

When we build a mathematical model of a system of causal influence, we will introduce names for variables and specify their possible values. Just as we distinguish numbers from numerals, so too should we distinguish the causal relata—*variables*—from their mathematical representation—*variable names*. The variable *whether I throw the ball* is a class of classes of spacetime regions. But we can denote this variable with a label—for instance, 'B'. 'B' is a variable name. We could say that the possible values for the variable name 'B' are 0 and 1, with $B = 0$ corresponding to me not throwing the ball and $B = 1$ corresponding to me throwing the ball. Within the mathematical model, '$B = *$' will not be a well-formed expression. That's because the mathematical model will presuppose that all of the variables of interest take on some value or other. A *signature*, $\mathcal{S}$, gives us a name for every variable and specifies what its possible values are. It additionally tells us which variables are exogenous and which are endogenous (a distinction I will introduce below). Formally, a signature $\mathcal{S}$ is a triple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variable names, $\mathcal{V}$ a set of endogenous variable names, and $\mathcal{R}$ is a function from the variable names in $\mathcal{U} \cup \mathcal{V}$ to their potential values. (From here on out, I won't bother explicitly distinguishing variables from variable names. When I am talking about the labels in a mathematical model, I mean 'variable name'; when I am talking about the causal relata out in the world, I mean 'variable'.)

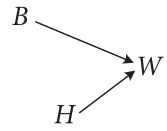
Relations of causal influence can be represented with a system of *structural equations*. For instance, suppose that you offer me a bet on whether a flipped coin will land heads. If I bet and the coin lands heads, then I get \$1. If I bet and the coin lands tails, then I lose \$1. If I don't bet, then I get nothing. Let 'B' represent the variable *whether I bet*. Say that B takes on the value 1 if I take the bet and it takes on the value 0 if I reject the bet. Likewise, let 'H' represent the variable *whether the coin lands heads*. It is 1 if the coin lands heads and -1 if it lands tails. And let 'W' name the variable *how much I win*. It is 1 if I win \$1, -1 if I lose \$1, and 0 if I neither win nor lose. Then, the following system of *structural equations* says that *how much I win* is causally influenced both by *whether I bet* and by *whether the coin lands heads*.

$$W := B \cdot H$$

This system of equations doesn't just tell me that W is causally influenced by B and H. It additionally tells me *how* B and H causally influence W. If $B = 0$, then B causally determines that $W = 0$.

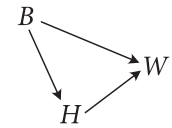
If $B = 1$ and $H = 1$, then B and H causally determine that $W = 1$. And if $B = 1$ and $H = -1$, then B and H causally determine that $W = -1$.

A structural equation is asymmetric. B and H causally influence W, but W does not causally influence either B or H. Given the system of equations $W := B \cdot H$, we may produce the following *directed graph* to illustrate the pathways along which the variables causally influence each other.



I said that $W := B \cdot H$ is a *system* of structural equations. It is a system with a single equation. We should distinguish the *system* of structural equations $W := B \cdot H$ from the structural equation $W := B \cdot H$. The latter could appear in the following system of structural equations:

$$W := B \cdot H$$
$$H := B$$



Whereas the *system* of structural equations $W := B \cdot H$ tells us that B and H are causally independent—neither causally influences the other—the structural equation $W := B \cdot H$ does not. It is consistent with B causally influencing H, H causally influencing B, or neither causally influencing the other. A *system* of equations effectively includes a *that's all* clause, telling us that the relations of causal influence the system describes are the *only* relations of causal influence which obtain between the variables it includes.[22]

In a system of structural equations, the variables which appear on the left-hand-side of an equation are called *endogenous*, and the ones which do not are called *exogenous*. Whether a variable is endogenous or exogenous is a property of which model we are looking at, and not the variable itself. I'll denote the set of exogenous variables in a model with '$\mathcal{U}$', and the set of endogenous variables in the model with '$\mathcal{V}$'. I'll take for granted that no variable lies causally downstream of itself. If that's so, then the determinism of the equations implies that, once we know which values the exogenous variables take on, we know which values every variable in the model takes on. So a model of a system of causal influence need only tell us which values each of the exogenous variables take on. We can specify which values the exogenous variables take on with an *assignment of values* to the exogenous variables in $\mathcal{U}$.

In general, given a set of variables $\mathbf{V}$, an *assignment of values*, $\mathbf{v}$, to the variables in $\mathbf{V}$ is a—perhaps partial—function from the variables $V \in \mathbf{V}$ to the values in $\mathcal{R}(V)$. Since the function $\mathbf{v}$ need not be total, it need not assign a value to every variable in $\mathbf{V}$. If $\mathbf{v}$ is an assignment of values to $\mathbf{V}$, then I'll write '$\mathbf{V} = \mathbf{v}$' for the claim that, for every $V \in \mathbf{V}$ in the domain of $\mathbf{v}$, $V = \mathbf{v}(V)$. That is, '$\mathbf{V} = \mathbf{v}$' says that, for each variable $V \in \mathbf{V}$ to which $\mathbf{v}$ assigns a value, V takes on the value which

[22] In some applications, we may want to impose a stronger requirement on a system of structural equations: that the variables are closed under common causal influence. That is: for all variables X, Y, Z: if X and Y are in the system and Z causally influences both X and Y, then Z is also included in the system. (This closure condition is often called *causal sufficiency*—see Spirtes et al. (2000) and Hausman & Woodward (1999), for instance.) Common causes could make a difference to the evaluation of backtracking counterfactuals, but they won't make any difference to the evaluation of causal counterfactuals. So I won't be assuming causal sufficiency here.

v assigns it. I'll call a total assignment of values to the exogenous variables in $\mathcal{U}$ an *exogenous assignment*.

What I will here call a *causal model*, $\mathcal{M}$, is a triple containing a *signature*, $\mathcal{S}$, a system of structural equations, $\mathcal{E}$, and an exogenous assignment, **u**. Or, equivalently, a causal model is a 5-tuple of a set of exogenous variables, a set of endogenous variables, a specification of the variables' potential values, a system of structural equations, and an exogenous assignment, $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{R}, \mathcal{E}, \mathbf{u})$.

A causal model represents a system of causal influence. Ideally, we would want to be able to model non-deterministic systems of causal influence. This would require more complicated causal models, but it would not affect anything I have to say here about the causal influence between your choice, the laws, and the distant past. So I'll stick to the deterministic case in the interests of simplicity. In the deterministic case, causal influence between variables goes along with causal *determination* between variable *values*. Thus, according to the structural equation $Y := \phi(X)$, the variable X causally influences the variable Y, and a variable value $X = x$ causally determines the variable value $Y = \phi(x)$.

2.2 | Causal counterfactuals

Because a causal model contains an exogenous assignment and a system of structural equations, it tells us which value every variable in the model takes on. If the variable V takes on the value v in the model $\mathcal{M}$, then we may write '$\mathcal{M} \models V = v$', and say that $\mathcal{M}$ *validates* the formula '$V = v$'. This definition of validation may be extended to Boolean combinations of variable values in the usual way.

Because causal models explicitly represent systems of causal influence, we can additionally say whether a model validates a causal counterfactual conditional. Suppose we have an *antecedent* variable, A, and a *consequent* variable C. And we wish to know whether, were A to take on the value a, C would take on the value c, $A = a > C = c$. In a causal counterfactual, we hold fixed factors which are not causally downstream of the antecedent, and we allow to swing free factors which are causally downstream of the antecedent. Within a causal model, we can achieve this by removing A's structural equation, effectively severing any causal influence between A and its causal parents and 'exogenising' the variable A. Then, we may solve for the values of the other variables in the model as before. If it turns out that $C = c$ in this minimally altered model, then the counterfactual $A = a > C = c$ was validated by the original model.

More carefully, given a causal model $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{R}, \mathcal{E}, \mathbf{u})$, let us define the *minimally altered* model in which A takes on the value a, $\mathcal{M}[A \rightarrow a]$, as follows. If $A = a$, then $\mathcal{M}[A \rightarrow a]$ is just $\mathcal{M}$ itself.[23] If $A \neq a$ and A is exogenous, then $\mathcal{M}[A \rightarrow a]$ is just $\mathcal{M}$, with the exogenous assignment **u** altered to assign the value a to A. The most interesting case is when $A \neq a$ and A is endogenous (though this case won't actually be relevant to our discussion here). If A is endogenous and $A \neq a$, then $\mathcal{M}[A \rightarrow a]$ is the model you get by moving A from the endogenous to the exogenous variable set, removing A's structural equation (the one with A on the left-hand-side) from the system of equations, and adding $A = a$ to the exogenous assignment. Iff the minimally altered model $\mathcal{M}[A \rightarrow a]$ validates '$C = c$', the original model $\mathcal{M}$ validates the causal counterfactual

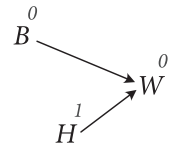
[23] Without this clause, causal counterfactuals will not satisfy *modus ponens*. See Briggs (2012).

'$A = a > C = c$'.

$$(>_{\mathcal{M}}) \qquad \mathcal{M} \vDash A = a > C = c \iff \mathcal{M}[A \to a] \vDash C = c$$

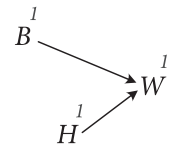
For illustration, consider the following causal model, which we can call '$\mathcal{M}_{coin}$':

$$\begin{aligned} B &= 0 \\ H &= 1 \\ W &:= B \cdot H \end{aligned}$$



This model contains the exogenous variable set $\mathcal{U} = \{B, H\}$, and the endogenous variable set $\mathcal{V} = \{W\}$. The range of B, $\mathcal{R}(B)$, is $\{1, 0\}$. The range of H, $\mathcal{R}(H) = \{1, -1\}$. And the range of W, $\mathcal{R}(W) = \{-1, 0, 1\}$. It contains the system of structural equations $\mathcal{E} = \{W := B \cdot H\}$. And the exogenous assignment maps B to 0 and H to 1. In $\mathcal{M}_{coin}$, the causal counterfactual '$B = 1 > W = 1$' ('had you taken the bet, you would have won') is true. For consider the minimally altered model $\mathcal{M}_{coin}[B \to 1]$:

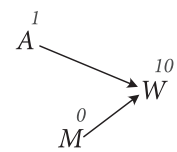
$$\begin{aligned} B &= 1 \\ H &= 1 \\ W &:= B \cdot H \end{aligned}$$



In this model, the consequent '$W = 1$' is true. The reason this counterfactual comes out true is that there is no causal influence from B to H. So, when we consider what would have happened, had you taken the bet, we hold fixed the actual value of H.

Or consider $\mathcal{M}_{miracle}$, which models the decision you face in **Betting on a miracle**. This model contains the variable A, for which bet you choose. If you choose bet a, then $A = 1$. If you choose bet b, then $A = 0$. It also contains a variable, M, for whether there is a miracle. If there is a miracle, then $M = 1$; and if there is not, then $M = 0$. Finally, there is a variable, W, for how much money you win. W can take on any value in $\{0, 1, 10, 11\}$, and its value is equal to the number of dollars you win. Suppose you actually take bet a, and there is no miracle.

$$\begin{aligned} A &= 1 \\ M &= 0 \\ W &:= 10 \cdot A \cdot \overline{M} + 11 \cdot \overline{A} \cdot M + \overline{A} \cdot \overline{M} \end{aligned}$$

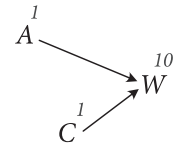


(In this structural equation, '$\overline{X}$' is the function $1 - X$.) In $\mathcal{M}_{miracle}$, the causal counterfactual $A = 0 > M = 0$ ('had you not chosen a, there would not have been a miracle') is true. For, in the minimally altered model $\mathcal{M}_{miracle}[A \to 0]$, the value of M remains 0. The reason the counterfactual comes out true in $\mathcal{M}_{miracle}$ is that there's no causal influence between A and M. So, when we consider what would have happened, had you not chosen a, we hold fixed whether there was a miracle.

Finally, consider $\mathcal{M}_{past}$, which models the decision you face in **Betting on the Past**. Like $M_{miracle}$, this model contains the variables A and W, with the same values and the same

interpretations. It also contains the variable C, for whether the initial conditions are c. If the initial conditions are c, then $C = 1$; and if they are not c, then $C = 0$. Suppose you actually take the bet a, and the initial conditions are c.

$$A = 1$$
$$C = 1$$
$$W := 10 \cdot A \cdot C + 11 \cdot \overline{A} \cdot \overline{C} + \overline{A} \cdot C$$



In $\mathcal{M}_{past}$, the counterfactual $A = 0 > C = 1$ ('had you not chosen a, the initial conditions would have been c') is true. For, in the minimally altered model $\mathcal{M}_{past}[A \rightarrow 0]$, the value of C remains 1. The reason the counterfactual comes out true in $\mathcal{M}_{past}$ is that there's no causal influence between A and C. So, when we consider what would have happened, had you not chosen a, we hold fixed the initial conditions.

It is one thing to write down these causal models and show that they validate the counterfactuals (B1) and (B2). It is another thing to show that these are the *correct* models to be using to evaluate the counterfactuals. Take any counterfactual you wish—'had I not cut my toenails on November 8th, 2016, Trump wouldn't have won', for instance. It's completely trivial to write down a causal model according to which this counterfactual is true. Just use the variable C, for whether I cut my toenails, and T, for whether Trump wins, and include the structural equation $T := C$. Simply writing down this model isn't enough to show that whether Trump won counterfactually depends upon whether I cut my toenails. And likewise, writing down the models $\mathcal{M}_{miracle}$ and $\mathcal{M}_{past}$ above isn't enough to show that the counterfactuals (B1) and (B2) are true. We could after all just as easily have written down models according to which whether you choose a causally influences whether there's a miracle or whether the initial conditions are c.

$(>_{\mathcal{M}})$ tells us what it is for a model to validate a causal counterfactual conditional. On its own, that does not tell us what it is for a causal counterfactual conditional to be true or false. I will take it for granted here that a causal counterfactual conditional is *true* if it is validated by a causal model which adequately represents the relations of causal influence out in the world—or, for the sake of brevity: the conditional is true if it is validated by a *correct* causal model. Likewise, the counterfactual is false if its negation is validated by a correct causal model. (If there is no causal model which validates either the counterfactual or its negation, then I say nothing about whether the counterfactual is true or false.)

$$(>) \qquad \begin{array}{l} \exists \mathcal{M} : \mathcal{M} \text{ is correct } \wedge \mathcal{M}[A \rightarrow a] \models C = c \Rightarrow A = a > C = c \\ \exists \mathcal{M} : \mathcal{M} \text{ is correct } \wedge \mathcal{M}[A \rightarrow a] \models C \neq c \Rightarrow A = a \not> C = c \end{array}$$

Then, if the models $\mathcal{M}_{miracle}$ and $\mathcal{M}_{past}$ are going to offer a satisfying resolution of the puzzle from §1.2, we must be told something general about when a causal model is correct. And we must be given reason to think that the models $\mathcal{M}_{miracle}$ and $\mathcal{M}_{past}$ are correct. This is the task I will take up in §§3 and 4 below. In §3, I will sketch a theory of causal influence—a theory of when a causal model is correct. In §4, I will explain how this theory tells us that, in the relevant decisions, $\mathcal{M}_{miracle}$ and $\mathcal{M}_{past}$ correctly describe the causal influence your choices have over whether there's a miracle and whether the initial conditions are c, respectively.

3 | A THEORY OF CAUSAL INFLUENCE

Standard semantics for counterfactuals utilise a *selection function* on a space of possible worlds. In this framework, each proposition is taken to be a set of possible worlds, a proposition is *true at* a world iff the world is contained within the proposition, and one proposition, A, entails another, B, iff $A \subseteq B$. A *selection function*, s, is a function from a proposition, A, and a possible world, w, to a proposition, $s(A, w)$. The standard semantics then say that $A > C$ is true at a world w iff $s(A, w)$ entails C.[24]

$$(>_s) \qquad w \in A > C \quad \Longleftrightarrow \quad s(A, w) \subseteq C$$

For causal counterfactuals, I reject $(>_s)$. In its place, I accept $(>)$. But I will nonetheless utilise the framework of the standard semantics to say when a causal model correctly represents a system of causal influence. That is: I will appeal to a space of possible worlds and a selection function to explain what it takes for a causal model to be correct.

For illustration, consider two variables, X and Y, with two possible values, 1 and 0. Then, propositions like $X = 0$ and $Y = 1$ will correspond to sets of possible worlds—the set of possible worlds in which those variables take on those values.[25] Now, consider the structural equation $Y := X$. I will say that, if this structural equation is correct at w, then

$$s(X = 0, w) \subseteq Y = 0 \quad \text{and} \quad s(X = 1, w) \subseteq Y = 1$$

Think of $s(A, w)$ as a set of A-worlds which are not too different from w. Then, I will say that, at w, X causally influences Y in the way described by the equation $Y := X$ only if (a) the set of $X = 0$ worlds which are not too different from w are all worlds at which $Y = 0$; and (b) the set of $X = 1$ worlds which are not too different from w are all worlds at which $Y = 1$. (When I say that the worlds are "not too different", I mean to appeal to your intuitive standards of similarity, applied to the time of the antecedent. Of course, minor differences at one time can balloon into large differences at a later time. This famously led to trouble for Lewis's interpretation of $s(A, w)$ as the set of A-worlds not too different from w *tout court*.[26] Lewis attempted to deal with the problem by introducing stipulative standards of similarity, but his attempts were not successful.[27] From my perspective, it is better to rely on our intuitive standards of similarity, but restrict the kinds of similarities which matter. Differences at the time of the antecedent are relevant, but even large differences at other times are not. There is more to be said here, but fortunately, not much will hang upon the particulars of how we understand the selection function. Whenever the details become relevant, I'll explicitly discuss them.)

Below, I will say something slightly more general about the relationship between a causal model and a selection function. In §4, I will use this general theory of causal influence to explain why the causal models $\mathcal{M}_{miracle}$ and $\mathcal{M}_{past}$ from §2 are correct. Because these models contain only a

[24] See Stalnaker (1968) and Lewis (1973). My presentation here rejects Stalnaker's uniqueness assumption, but accepts his limit assumption.

[25] I assume that, necessarily, a variable takes on a value in at most one spacetime region. Thus, for instance, $X = 0 \cap X = 1 = \varnothing$.

[26] See Lewis (1973), Bennett (1974), Fine (1975), Lewis (1979), and Bennett (2003, §75).

[27] See, for instance, Elga (2001) and Wasserman (2006).

single structural equation, I'll focus on this special case here. A more general treatment can be found in Gallow (2016).

Take a causal model containing the system of equations $\mathcal{E} = \{V := \phi(\mathbf{PA}(V))\}$. (By the way, $\mathbf{PA}(V)$ are V's causal *pa*rents—the variables which appear on the right-hand-side of V's structural equation—and '$\phi(\mathbf{PA}(V))$' is some function of all and only the variables in $\mathbf{PA}(V)$.) In order for this causal model to be correct, all of the variables appearing in $\mathbf{PA}(V) \cup \{V\}$ must be mereologically *distinct*; they must not overlap. This distinctness requirement is an important component of most theories of causation. For instance, take a counterfactual theory of causation, and consider the events *your playing cards* and *your playing poker*. If you hadn't played cards, you wouldn't have played poker. We should not conclude that your playing cards *caused* you to play poker. The connection between these events is *constitutive*, not causal. For this reason, careful counterfactual theories of causation stipulate that counterfactual dependence reveals causation only when the two events are mereologically distinct.[28] And for similar reasons, we should not allow a causal model to include variables which overlap. The mereology of variables is another topic for another occasion. But let me offer the following necessary (but *in*sufficient) condition on all of the variables in $\mathbf{PA}(V) \cup \{V\}$ being distinct: every assignment of values to these variables must be metaphysically possible. That is: for every assignment of values to the variables in $\mathbf{PA}(V) \cup \{V\}$, it must be possible that those variables take on those values. This is the condition which Woodward (2015) calls *independent fixability*.[29] (I'll have a bit more to say about this requirement, and why it's important, below.)

In addition, if the system of equations $\{V := \phi(\mathbf{PA}(V))\}$ is going to be correct, then the variables in $\mathbf{PA}(V)$ should causally influence V in the manner described by ϕ. And I will say that this is so iff, in all of the not too different possibilities in which we wiggle the values of the 'parent' variables, $\mathbf{PA}(V)$, the equality $V = \phi(\mathbf{PA}(V))$ continues to hold. That is: if the system $\{V := \phi(\mathbf{PA}(V))\}$ is correct at w, then the following must be true, for every assignment of values, $\mathbf{pa}$, to the variables in $\mathbf{PA}(V)$:

$$s(\mathbf{PA}(V) = \mathbf{pa}, w) \subseteq V = \phi(\mathbf{PA}(V))$$

That is: to check whether $V := \phi(\mathbf{PA}(V))$ is correct at w, you have to take every assignment of values to $\mathbf{PA}(V)$ and consider every world not too different from w in which that assignment is realised. For each such world, you must check that the value of V at that world equals the value to which ϕ maps the values of $\mathbf{PA}(V)$ at that world. This imposes a kind of *stability* requirement on the system of equations $\{V := \phi(\mathbf{PA}(V))\}$. The equation $V := \phi(\mathbf{PA}(V))$ must not only be *actually* true; it must also be that it *remains* true, no matter how we wiggle the values of the parent variables in $\mathbf{PA}(V)$.

As I emphasised in §2 above, the system of equations $\{V := \phi(\mathbf{PA}(V))\}$ doesn't just say that each $P \in \mathbf{PA}(V)$ causally influences V. It also says that none of the $P \in \mathbf{PA}(V)$ are causally influenced by any other variables in the model. Suppose that, for some 'parent' variable $P \in \mathbf{PA}(V)$, there is a set of variables, $\mathbf{Q}$, which includes at least one other variable from the model— i.e., $\mathbf{Q} \cap \mathbf{PA}(V) \neq \emptyset$—such that the variables in $\mathbf{Q}$ causally influence P. If that's so, then the system of equations $\{V := \phi(\mathbf{PA}(V))\}$ is not correct. For, if that's so, then there is causal influence between some of the variables in the model, but the model does not *tell* us about that causal influence. Some collection of variables $\mathbf{Q}$ causally influences P if there's some function ψ such that, in all of the not

[28] See the discussion in Lewis (1986a, 1986b).

[29] For further discussion of the mereology of variables, see Hoffmann-Kolss (2021).

too different possibilities in which we wiggle the values of **Q**, the equation $P = \psi(\mathbf{Q})$ continues to hold. So, if the system of equations $\{V := \phi(\mathbf{PA}(V))\}$ is going to be correct, then there cannot be a set of variables **Q** and a function ψ like that. (Of course, not just *any* function ψ is enough to reveal genuine causal influence. If ψ is a *constant* function of **Q**, this doesn't reveal any influence that **Q** has on P. In general, I think we should require that ψ be both a non-constant function and a function *of* every variable $Q \in \mathbf{Q}$. For instance, the function $\psi(X) = 1$ does not count, since it is constant, and $\psi(X, Y) = X + (Y - Y)$ does not count, since it is not a function of Y. Of course, all the same remarks apply to the function ϕ in the system of equations $\{V := \phi(\mathbf{PA}(V))\}$. It too must be a non-constant function of all of the 'parent' variables in $\mathbf{PA}(V)$.)

Putting these three requirements together, we get:

Causal Influence The system of equations $\mathcal{E} = \{V := \phi(\mathbf{PA}(V))\}$ is correct at a world w iff

(E1) all of the variables in $\mathbf{PA}(V) \cup \{V\}$ are distinct;

(E2) for every assignment of values to $\mathbf{PA}(V)$, $\mathbf{pa}$,

$$s(\mathbf{PA}(V) = \mathbf{pa}, w) \subseteq V = \phi(\mathbf{PA}(V))$$

and

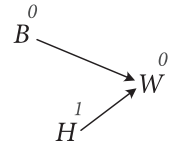
(E3) for every $P \in \mathbf{PA}(V)$, there is no set **Q** containing variables from $\mathbf{PA}(V)$ such that (a) all of the variables in $\mathbf{Q} \cup \{P\}$ are distinct, and (b) there's a non-constant function ψ of the variables in **Q** such that, for every assignment of values to **Q**, **q**,

$$s(\mathbf{Q} = \mathbf{q}, w) \subseteq P = \psi(\mathbf{Q})$$

Condition (E1) tells us that there are no (metaphysically) necessary connections between the variables' values. Condition (E2) tells us that the variables in $\mathbf{PA}(V)$ all causally influence V in the way described by the function ϕ. And condition (E3) tells us that none of the variables in $\mathbf{PA}(V)$ causally influence each other.

We can illustrate **Causal Influence** by showing how it can be used to vindicate the causal model $\mathcal{M}_{coin}$ in the decision described in §2 above.

$$\begin{aligned} B &= 0 \\ H &= 1 \\ W &:= B \cdot H \end{aligned}$$



Causal Influence tells us that, in order for $\mathcal{M}_{coin}$ to be correct, the variables B, H, and W must not overlap—in particular, they must be *independently fixable*. We must be careful here. In particular, we must understand the variable W in such a way that you winning \$1 ($W = 1$) does not entail that you took the bet. That is: we must understand the variable W in such a way that you could win \$1 without winning \$1 *off of this very bet*. In that case, every assignment of values to the variables will be metaphysically possible. And, moreover, the variables in $\{B, H, W\}$ will all be mereologically distinct. So condition (E1) is satisfied.

Condition (E2) says that, in every not too different possibility in which one of the assignments of values to $\{B, H\}$ is realised, the value of W must be equal to $B \cdot H$. Because there are 8 possible

assignments of values to $\{B, H\}$, this imposes 8 different constraints. Assuming that, for any A, $s(A, w) \subseteq A$, condition (E2) requires each of the following:

$$s(B = 0, w) \subseteq W = 0 \qquad s(B = 1, w) \subseteq W = H$$

$$s(H = -1, w) \subseteq W = -B \qquad s(H = 1, w) \subseteq W = B$$

$$s(B = 0 \wedge H = -1, w) \subseteq W = 0 \qquad s(B = 0 \wedge H = 1, w) \subseteq W = 0$$

$$s(B = 1 \wedge H = -1, w) \subseteq W = -1 \qquad s(B = 1 \wedge H = 1, w) \subseteq W = 1$$

(Here, w is the world we are modelling; it is the world in which you refuse the bet and the coin lands heads.) Assuming that the betting arrangement remains intact at any world not too different from w at which B and H are assigned these values, each of these constraints should be satisfied. For illustration, take the first two constraints. The first says: any possibility not too different from w at which you refuse the bet must be one at which you neither win nor lose any money. This constraint will be satisfied; for, if you refuse the bet, then it won't matter how the coin lands, you'll neither gain or lose any money. The second says: any possibility not too different from w at which you take the bet must be one at which the value of W is equal to the value of H. This constraint, too, will be satisfied. Either the coin will land tails, $H = -1$, and you will lose \$1, or else the coin will land heads, $H = 1$, and you will win \$1. Either way, the value of W will equal the value of H.

Finally, condition (E3) requires that neither B nor H causally influence the other. If we suppose that $s(B = 1, w)$ contains both possibilities at which the coin lands heads and possibilities at which the coin lands tails, then B will not *on its own* causally influence H. For $s(B = 1, w)$ does entail that H is any function of B—the value of H varies while the value of B is held fixed. So H is not causally influenced by B. It is also natural to suppose that both $s(H = 1, w)$ and $s(H = -1, w)$ contain only worlds at which you (still) refuse the bet. Making the coin land heads or tails does not require us to change anything about your preceding decision. If so, then B will not be causally influenced by H. This doesn't establish that B doesn't causally influence H *in concert with* some other variables, but no candidates spring to mind. There are of course variables which causally influence whether the coin lands heads (the coin's precise initial upward and angular velocities, e.g.) but these variables causally influence whether the coin lands heads *on their own*—we do not need the extra information of whether you took the bet or not. So I will take it for granted that condition (E3) is satisfied, though I do not pretend to have conclusively demonstrated this.

So **Causal Influence** tells us that, in this decision, $\mathcal{M}_{coin}$ is correct. Then, $(>)$ tells us that the causal counterfactual "had you taken the bet, you would have won" $(B = 1 > W = 1)$ is true. This is noteworthy for three reasons. Firstly, the counterfactual appears true. Secondly, counterfactuals like these have important theoretical roles to play elsewhere. Suppose, for instance, that I talked you out of taking the bet. Then, it seems that I prevented you from winning \$1. If we accept a counterfactual theory of causation, then we'll want the causal counterfactual "had I not talked you out of taking the bet, you would have won \$1" to be true.[30] Thirdly, holding fixed our assumptions about the selection function, the semantics $(>_s)$ will tell us that $B = 1 > W = 1$ is false. For we

[30] The truth of a counterfactual like this isn't in general needed for the corresponding claim about prevention to be true. It could be, for instance, that, had I not talked you out of the bet, someone else would have. But if there's no funny business like that going on, then we should expect the truth of the prevention claim to go along with the truth of the causal counterfactual claim.

assumed that $s(B = 1, w)$ contains both worlds where the coin lands heads (and, therefore, you win) and worlds where the coin lands tails (and, therefore, you lose). Then, $s(B = 1, w) \not\subseteq W = 1$, so according to $(>_s)$, $B = 1 \not> W = 1$.[31] Of course, we could always reject one of our assumptions about the selection function. Standard ways of doing this require us to characterise the selection function in terms of causal influence.[32] If we were to then characterise causal influence in terms of the selection function, our theory would be circular—not viciously circular, in my view, but circular nonetheless.

A non-circular theory of causal counterfactuals would be preferable, other things being equal. For a non-circular theory allows us to explain things which a circular theory does not. For instance, it allows us to explain why the outcome of the coin flip is not causally influenced by whether you take the bet. So I take it to be a benefit of the theory I've sketched here that—without any assumptions about causal influence—it predicts that the causal counterfactual "had you taken the bet, you would have won" is true. This prediction gives us some reason to accept the theory, quite independent of the fact that it vindicates both (B1) and (B2).

With this theory of causal influence in place, let me return to the requirement of mereological distinctness, (E1). It's worth considering what would happen if we dropped this requirement. Suppose you will win money iff a rolled die lands on 5, the die is rolled, it lands on 5, and you win. Without (E1), we could model this situation with three variables: H, for whether the die lands *h*igh, O, for whether the die lands *o*dd, and W, for whether you *w*in. And we could use the structural equation $W := H \wedge O$. The variables H and O are not mereologically distinct, so condition (E1) tells us that this system of equations is not correct.[33] But the system of equations satisfies condition (E2). Even if there's some larger system of equations in which $W := H \wedge O$ is embedded, so that either H is causally downstream of O or O is causally downstream of H, so long as the model is acyclic, we will have to accept either 'had the die not landed high, it still would have landed odd' or 'had the die not landed odd, it still would have landed high'. But both of these counterfactuals appear false.

There are interesting questions about how—or whether—a semantics like $(>)$ should be extended to handle counterfactuals involving overlapping variables. Fortunately, we won't have to open that can of worms in order to show that the models $\mathcal{M}_{miracle}$ and $\mathcal{M}_{past}$ are correct in the relevant decisions, since those models don't involve overlapping variables. However, the question of how to handle causal counterfactuals involving overlapping variables will be relevant to a decision I'll discuss in §5.

4 | CAUSAL COUNTERFACTUALS WITHOUT MIRACLES OR BACKTRACKING

In this section, I will explain how the theory adumbrated in §3 above can be used to show that $\mathcal{M}_{miracle}$ and $\mathcal{M}_{past}$ are correct in the relevant decisions. This will show that, in those decisions, the causal counterfactuals 'had you not chosen a, there wouldn't have been a miracle' (B1) and 'had you not chosen a, the initial conditions would have been c' (B2) are both true.

[31] Sidney Morgenbesser raised this as an objection to $(>_s)$. See Slote (1978).

[32] See, for instance, the proposals in Bennett (2003, ch. 15), Edgington (2004), Schaffer (2004), and Kment (2006).

[33] Incidentally, this example—taken from Hoffmann-Kolss (2021)—shows us why independent fixability is not sufficient for mereological distinctness. But Lewis (1986a)'s theory of event mereology, naturally generalised to apply to variables, will tell us that H and O overlap.

In order to get the result that $\mathcal{M}_{miracle}$ and $\mathcal{M}_{past}$ are correct, I will have to make an additional stipulation about the selection function I'm using in my theory of causal influence. I'll explain this additional stipulation in §4.1. Then, in §4.2, I will explain why the models $\mathcal{M}_{miracle}$ and $\mathcal{M}_{past}$ are correct.

4.1 | Strong centring

I'm going to take for granted that the selection function we're using in the theory of causal influence from §3 satisfies some standard structural constraints like: 1) for all A, $s(A, w) \subseteq A$; and 2) for all A, B, if $A \subseteq B$ then $s(B, w) \cap A \subseteq s(A, w)$. Importantly, however, I won't impose the following structural constraint, known as *strong centring*.

Strong Centring If $w \in A$, then $s(A, w) = \{w\}$.

Strong centring says that, if A is already true at w, then w itself is the only world not too different from w at which A is true. In other words: if it's possible for A to be true without things being any different than they are at w, then any difference from w whatsoever is *too* different from w. Given the standard semantics, strong centring corresponds to the principle of conjunction conditionalisation (CC), which allows you to infer $A > C$ from $A \wedge C$.[34] However, if we reject the standard semantics, there needn't be any relationship between strong centring and CC. Indeed, CC follows from the causal-modelling semantics ($>$) which I provided in the previous section.[35] So, if we accept ($>$), there need not be any connection between strong centring and CC.

To appreciate why I do not want to impose strong centring, consider the variables J and D, which represent *whether Jesus of Nazareth is born* and *whether the Defenestration of Prague occurs*, respectively. $J = 1$ if Jesus is born, and $J = 0$ if he is not born. $D = 1$ if the Defenestration of Prague happens and $D = 0$ if it does not. And consider the structural equation $D := J$. According to this equation,

(F1) Jesus not being born causally determines the Defenestration to not happen,

$$s(J = 0, w) \subseteq D = 0$$

and

(F2) Jesus being born causally determines the Defenestration to happen,

$$s(J = 1, w) \subseteq D = 1$$

(I'm using 'w' for the actual world.) It appears that (F1) is true. After all, had Jesus never existed, neither would the Catholic Church have existed; and without the Catholic Church, there would be no Protestant Reformation, nor the Bohemian religious disputes which precipitated the Defenestration of Prague. Any not too different possible world in which Jesus is not born is a world too different for the Defenestration of Prague to occur. Now, if I were to impose strong centring, then (F2), $s(J = 1, w) \subseteq D = 1$, would be automatic. For w itself is a world at which Jesus is born. Strong

[34] See Walters & Williams (2013) for an argument for conjunction conditionalisation.

[35] Here, it is important that I defined $\mathcal{M}[A \rightarrow a]$ to be $\mathcal{M}$ itself, if A already takes on the value a in $\mathcal{M}$. See Briggs (2012).

centring would then tell us that $s(J = 1, w) = \{w\}$. And since w is also a world at which the Defenestration occurs, $\{w\} \subseteq D = 1$. But the structural equation $D := J$ appears false. Even if there is a convoluted chain of causal influence connecting these two variables, the value of J is not *directly causally sufficient* for the value of D, in the way that structural equation $D := J$ requires.

While I want my selection function to validate (F1), I want it to falsify (F2). Then, since both (F1) and (F2) are needed for the structural equation $D := J$ to be correct, my theory will say that the equation is not correct. The way I will falsify (F2) is by using a selection function which is not strongly centred. Jesus could easily have been born in a variety of different ways. I'll want to include each of these easy ways for Jesus to be born in the set $s(J = 1, w)$. So, while I'll want w to be a member of $s(J = 1, w)$, I won't want it to be *the only* member. The world is chaotic, and minor variations in the manner of Jesus's birth make for larger differences in the course of his life and the lives of those around him, which lead to even larger differences in the course of human history hundreds of years down the line. Had Jesus been born with a birthmark or a cleft palate, his childhood and psychological development could easily have been vastly different; he could easily fail to become a religious leader, and even if he had become a religious leader, the reception of his teachings could easily have been vastly different. Minor differences snowball quickly enough that many, many of these easy possibilities are ones in which the Christian religion is never founded, or never adopted as a state religion by Constantine. Due to the extreme sensitivity of genetics on initial conditions—minor variations in the time and manner of copulation make for differences in which sperm fertilises which egg—a great many of them are possibilities in which none of the people who actually lived in the sixteenth century ever even existed. Without either a Catholic Church or a Martin Luther, these are possibilities in which the Bohemian religious disputes which precipitated the Defenestration of Prague never happened. So, as I want to understand the selection function, we won't have $s(J = 1, w) \not\subseteq D = 1$, and (F2) will be false. So I'll say that the structural equation $D := J$ is not correct.

To be clear: this is a stipulation, not a substantive assumption about the semantics of English-language counterfactuals. The selection function I'm using is just a function from worlds and propositions to propositions. It is a theoretical gadget, introduced to play a certain role in my theory of causal influence. We can specify how to understand a gadget like this by saying things like 'consider the possibilities which are not-too-different from w at the time of the antecedent, and at which A is true', or 'consider all the ways of locally wiggling the variable V so that it takes on the value v at the relevant time, as far as possible leaving everything else at that time unchanged, and then time-evolving everything into the future/past according to the laws of nature—holding fixed the universe's low-entropy initial conditions'.[36] And these specifications don't rely upon English-language counterfactuals. If we use this gadget to explain what it takes for a causal model to be correct in the way I proposed in §3, then we have good reason to not impose strong centring. For we should not want to say that all past historical events causally determine all events in the far enough future. So we should distinguish X's value *causally determining* Y's from Y's value *sensitively depending upon* X's. And drawing this distinction requires us to attend to more than a single possibility in which X takes on its actual value.

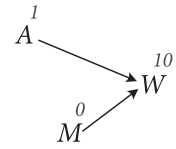
So I won't impose strong centring. However, I still will impose *weak* centring, which says that, if w is a world at which A is true, then w is *among* the A-worlds which are not too different from w. That is: if $w \in A$, then $w \in s(A, w)$. Weak centring will play an important role in my discussion below.

[36] Cf. Kutach (2002), Albert (2003), Loewer (2007), Hall (2007), Maudlin (2007), and Paul & Hall (2013), among others.

4.2 | Betting on a miracle

Consider again the decision **Betting on a miracle** from §1.3. Recall: in this decision, you must choose between two bets: a and b. Bet a pays out \$10 if there's not a miracle and nothing if there is; bet b pays out \$11 if there's a miracle and \$1 if there's not. At the actual world, w, you choose a and there's no miracle. In §2.2, I modelled this decision with the following system of equations.

$$\begin{aligned} A &= 1 \\ M &= 0 \\ W &:= 10 \cdot A \cdot \overline{M} + 11 \cdot \overline{A} \cdot M + \overline{A} \cdot \overline{M} \end{aligned}$$



According to **Causal Influence**, this system of equations is correct iff (E1) all of its variables are distinct, (E2) in all of the not too different possibilities in which we wiggle the values of A and M, the equation $W = 10 \cdot A \cdot \overline{M} + 11 \cdot \overline{A} \cdot M + \overline{A} \cdot \overline{M}$ continues to hold, and (E3) neither A nor M causally influences each other.

(E1) is satisfied. To appreciate this, notice that every assignment of values to the variables in $\{A, M, W\}$ is possible, so the variables are *independently fixable*.[37] As with the model $\mathcal{M}_{coin}$, we must exercise some caution here. In particular, we must understand the variable W in such a way that you could win \$1, \$10, or \$11 without winning it *off of this very bet*. (Otherwise, $W = 11$ will metaphysically necessitate that $A = 0$.) However, if we understand W in this way, then condition (E1) will be satisfied.

Condition (E2) will be satisfied so long as, in every not too different possibility in which one of the assignments of values to $\{A, M\}$ is realised, the value of W continues to be $10 \cdot A \cdot \overline{M} + 11 \cdot \overline{A} \cdot M + \overline{A} \cdot \overline{M}$. Because there are 8 assignments of values to $\{A, M\}$, this imposes 8 different constraints:

$$s(A = 0, w) \subseteq W = 11 \cdot M + \overline{M} \qquad s(A = 1, w) \subseteq W = 10 \cdot \overline{M}$$

$$s(M = 0, w) \subseteq W = 10 \cdot A + \overline{A} \qquad s(M = 1, w) \subseteq W = 11 \cdot \overline{A}$$

$$s(A = 0 \wedge M = 0, w) \subseteq W = 1 \qquad s(A = 0 \wedge M = 1, w) \subseteq W = 11$$

$$s(A = 1 \wedge M = 0, w) \subseteq W = 10 \qquad s(A = 1 \wedge M = 1, w) \subseteq W = 0$$

(Here, 'w' is the world at which you choose a and there's no miracle.)

There are many choices to be made about the worlds returned by the selection function. For instance, we could take the Lewisian route of saying that $s(A, w)$ contains worlds with the same past as w, in which a miracle occurs just before the time of the antecedent. Or we could instead side with authors like Dorr and say that $s(A, w)$ contains worlds in which there is no miracle, and so the past is ever-so-slightly different at a microphysical level. Call the first understanding of the selection function 'miraculous', and call the second a 'backtracking' understanding. For our purposes, it won't matter whether we adopt a backtracking or a miraculous understanding of the selection function.

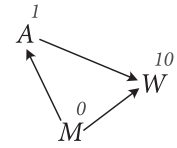
[37] Independent fixability is not sufficient for the variables to be distinct, but these variables also satisfy any stronger criteria we might reasonably want to impose.

So long as our betting arrangement remains intact at the not too different worlds at which the variables A and M are assigned these values, each of the eight constraints imposed by (E2) will be satisfied. For illustration, just take the first constraint. Amongst the not too different worlds where I take bet b, how much I win varies as a function of whether there's a miracle. If we have a miraculous understanding of the selection function, then all of the worlds in $s(A = 0, w)$ will contain a miracle. If we have a backtracking understanding, then none of them will. But, either way, the equality $W = 11 \cdot M + \overline{M}$ will hold. For, in the miraculous worlds, I'll win \$11; and in the non-miraculous worlds, I'll win \$1.

When we consider the other 7 constraints, we should guard against a potential confusion. On a miraculous understanding of the selection function, $s(A, w)$ *generally* takes us to worlds in which there's been a miracle to bring about A. However, if our antecedent explicitly stipulates that there is no miracle, $M = 0$, then the set of not too different worlds in which there's no miracle, $s(M = 0, w)$, must not include any miraculous worlds. If $s(M = 0, w)$ contains worlds other than w itself—as I argued it should in §4.1—then some of these worlds will be backtracking worlds at which the past is ever-so-slightly different. For, given that the laws are actually deterministic, *every* non-actual world is either a miraculous world or a backtracking world. And antecedents which explicitly stipulate that there is no miracle will forbid us from considering the miraculous worlds. So, if we must consider some non-actual worlds, we must consider some backtracking ones. This is consistent with the selection function *generally* delivering miraculous worlds. Likewise, on a backtracking understanding of the selection function, it *generally* takes us to non-miraculous worlds. However, if our antecedent explicitly stipulates that there is a miracle, $M = 1$, then the set of not too different worlds in which there's a miracle must consist of miraculous worlds. This is consistent with the selection function *generally* delivering non-miraculous worlds.

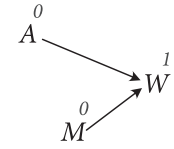
Condition (E3) requires that neither A nor M causally influence the other. However, for our purposes here, the only relevant requirement is that A not causally influence M. The reason is that, even if M causally influences A, the causal counterfactual $A = 0 > M = 0$ ('had you not taken a, there wouldn't have been a miracle') will still be true. That is, suppose that, in fact, a causal model like this one is correct:

$$\begin{aligned} M &= 0 \\ A &:= \overline{M} \\ W &:= 10 \cdot A \cdot \overline{M} + 11 \cdot \overline{A} \cdot M + \overline{A} \cdot \overline{M} \end{aligned}$$



If we begin with this model, then the minimally altered model in which A takes on the value 0 is shown below.

$$\begin{aligned} M &= 0 \\ A &= 0 \\ W &:= 10 \cdot A \cdot \overline{M} + 11 \cdot \overline{A} \cdot M + \overline{A} \cdot \overline{M} \end{aligned}$$



And this is precisely the same as the minimally altered model we get if we begin with a model in which M does not causally influence A. Since $M = 0$ in this minimally altered model, the causal counterfactual $A = 0 > M = 0$ will be true, whether or not M causally influences A.

Of course, if A causally influences M, this counterfactual needn't be true. So let us show that it does not. Again, it won't matter whether we have a miraculous understanding of the selection function or a backtracking understanding. Start with the miraculous understanding and suppose—for *reductio*—that there's some set of variables, $\mathbf{Q}$, which includes A and which is such that, for some non-constant function, ψ, of the variables in $\mathbf{Q}$, the equality $M = \psi(\mathbf{Q})$ is true in all of the not too different worlds in which we wiggle the values of some of the variables in $\mathbf{Q}$. That is: suppose that, for every assignment of values $\mathbf{q}$ to $\mathbf{Q}$, $s(\mathbf{Q} = \mathbf{q}, w) \subseteq M = \psi(\mathbf{Q})$. Consider the actual assignment of values, $\mathbf{q}_w$. By weak centring, $s(\mathbf{Q} = \mathbf{q}_w, w)$ contains w. And by stipulation, at w, there is no miracle, $M = 0$. So we have that $\psi(\mathbf{q}_w) = 0$. But since $s(\mathbf{Q} = \mathbf{q}_w, w)$ must contain some non-actual worlds (as I argued in §4.1), it must *also* contain some worlds at which there *is* a miracle (given the miraculous understanding of the selection function). So we have that $\psi(\mathbf{q}_w) = 1$. Contradiction. So if we adopt a miraculous understanding of the selection function, then there is no variable set containing A which causally influences M.

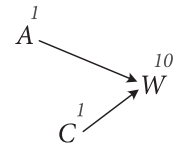
Next, take a backtracking understanding of the selection function. Suppose—for *reductio*—that there's a set of variables, $\mathbf{Q}$ which contains A and is such that $\mathbf{Q} \cup \{M\}$ are all distinct. Also suppose that there's a function ψ such that $M = \psi(\mathbf{Q})$ is true in all of the not too different possibilities in which the values of $\mathbf{Q}$ are wiggled. Since the selection function backtracks, $s(\mathbf{Q} = \mathbf{q}, w)$ will only contain miracles if the assignment $\mathbf{Q} = \mathbf{q}$ requires them. But if the assignment $\mathbf{Q} = \mathbf{q}$ metaphysically necessitates that $M = 1$, then $\mathbf{Q}$ and M will not be independently fixable, and the variables in $\mathbf{Q} \cup \{M\}$ will not be distinct. Since, by hypothesis, the variables *are* distinct, none of the worlds in $s(\mathbf{Q} = \mathbf{q}, w)$ will contain miracles—for *any* assignment $\mathbf{q}$. So $s(\mathbf{Q} = \mathbf{q}, w) \subseteq M = 0$ for every assignment $\mathbf{q}$. But then, $\psi(\mathbf{q}) = 0$ for every assignment $\mathbf{q}$. So ψ is a constant function. Contradiction. So if we adopt a backtracking understanding of the selection function, then there is no variable set containing A which causally influences M.

Either way, then, A does not causally influence M. Assuming that M doesn't causally influence A—though, to reiterate, it doesn't ultimately matter whether this is so—condition (E3) is satisfied. So the model $\mathcal{M}_{miracle}$ is correct. And so, given the semantics ($>$), the causal counterfactual 'if you hadn't chosen a, there wouldn't have been a miracle' (B1) is true.

4.3 | Betting on the past

Recall the decision **Betting on the past** from §1.3. You must choose between bet a and bet b. Bet a pays out \$10 if the initial conditions are c and nothing if they're not. And bet b pays out \$1 if the initial conditions are c and \$11 if they're not. In fact, the initial conditions are c and you choose a. In §2.2, I modelled this decision with the following system of equations

$$\begin{aligned} A &= 1 \\ C &= 1 \\ W &:= 10 \cdot A \cdot C + 11 \cdot \overline{A} \cdot \overline{C} + \overline{A} \cdot C \end{aligned}$$



Causal Influence tells us that this system of equations is correct iff (E1) all of its variables are distinct, (E2) in all of the not too different worlds where the values of A and C are wiggled, the equation $W = 10 \cdot A \cdot C + 11 \cdot \overline{A} \cdot \overline{C} + \overline{A} \cdot C$ is true, and (E3) neither A nor C causally influences the other.

Bearing in mind the caveat about the interpretation of the variable W from §4.2 above, condition (E1) will be satisfied. Condition (E2) imposes the following 8 constraints:

$$s(A = 0, w) \subseteq W = 11 \cdot \overline{C} + C \qquad s(A = 1, w) \subseteq W = 10 \cdot C$$

$$s(C = 0, w) \subseteq W = 11 \cdot \overline{A} \qquad s(C = 1, w) \subseteq W = 10 \cdot A + \overline{A}$$

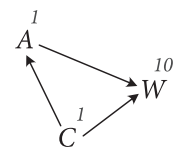
$$s(A = 0 \wedge C = 0, w) \subseteq W = 11 \qquad s(A = 0 \wedge C = 1, w) \subseteq W = 1$$

$$s(A = 1 \wedge C = 0, w) \subseteq W = 0 \qquad s(A = 1 \wedge C = 1, w) \subseteq W = 10$$

(Here, 'w' is the actual world, at which you choose a and the initial conditions are c.) So long as our betting arrangement remains intact at the not too different worlds at which the variables A and C are assigned these values, these 8 constraints should be satisfied, whether we have a miraculous or a backtracking understanding of the selection function. While it won't matter whether we think about the 'not too different' possibilities in terms of tiny miracles or in terms of minor changes to the past, it will matter very much which changes to the initial conditions we regard as *not too different*. Some ways of changing the initial conditions lead to large-scale macroscopic differences in the world at the time when you are offered the bet in the actual world. Others lead to minor, microscopic differences which only manifest in macroscopic differences after you are offered the bet. In the former kinds of possibilities, you may not be offered the bet at all, and the variable A may not take on a value. In the latter kinds of possibilities, the terms of the bet will remain the same, and the variables A and W will both take on values. I will take it for granted here that $s(C = 0, w)$ only contains the latter kinds of possibilities. And I'll assume likewise for $s(C = 1, w)$, $s(A = C = 0, w)$, and so on.

Condition (E3) requires that neither A nor C causally influence the other. However, just as in **Betting on a miracle**, it won't ultimately matter if C causally influences A. Even if a system of equations like this is correct,

$$\begin{aligned} C &= 1 \\ A &= C \\ W &:= 10 \cdot A \cdot C + 11 \cdot \overline{A} \cdot \overline{C} + \overline{A} \cdot C \end{aligned}$$



the causal counterfactual $A = 0 > C = 1$ ('had you not chosen a, the initial conditions would have been c') will still be true. So what matters is establishing that A doesn't causally influence C. For this purpose, it won't matter whether the selection function is miraculous or backtracking. Begin with the miraculous understanding. Suppose—for *reductio*—that there's some set of variables, $\mathbf{Q}$, including A, which is such that $\mathbf{Q} \cup \{C\}$ are distinct. Additionally suppose that there's a non-constant function ψ such that $C = \psi(\mathbf{Q})$ is true in all of the not too different worlds in which $\mathbf{Q} = \mathbf{q}$, for every assignment of values $\mathbf{q}$. Because $\mathbf{Q} \cup \{C\}$ are distinct, the variables in $\mathbf{Q}$ do not concern the state of the world at the initial conditions. Since the initial conditions are *initial*, the variables in $\mathbf{Q}$ must concern the state of the world at times *after* the initial conditions. Therefore, on the miraculous understanding, for every assignment $\mathbf{q}$, $s(\mathbf{Q} = \mathbf{q}, w)$ contains worlds with the same initial conditions as w. So it contains worlds at which $C = 1$. So, for every assignment $\mathbf{q}$, $\psi(\mathbf{q}) = 1$. But then ψ is a constant function. Contradiction. So if we adopt a miraculous understanding of the selection function, there is no variable set containing A which causally influences C.

Next, consider the backtracking understanding. Suppose—for *reductio*—that there's a set of variables, $\mathbf{Q}$, containing A, and a function ψ such that $C = \psi(\mathbf{Q})$ is true throughout the worlds in $s(\mathbf{Q} = \mathbf{q}, w)$, for every assignment $\mathbf{q}$. Consider the actual assignment $\mathbf{q}_w$. By weak centring, $s(\mathbf{Q} = \mathbf{q}_w, w)$ contains w. And by stipulation, at w the initial conditions are c. So $\psi(\mathbf{q}_w) = 1$. But since $s(\mathbf{Q} = \mathbf{q}_w, w)$ must contain some non-actual worlds (as I argued in §4.1), it must *also* contain some worlds at which the initial conditions are not c (given the backtracking understanding of the selection function). So we have that $\psi(\mathbf{q}_w) = 0$. Contradiction. So if we adopt a backtracking understanding of the selection function, then there is no variable set containing A which causally influences C.

Either way, then, A does not causally influence C. Assuming that C doesn't causally influence A—though, again, this doesn't ultimately matter—condition (E3) is satisfied. So the model $\mathcal{M}_{past}$ is correct. And so, given the semantics ($>$), the causal counterfactual 'if you hadn't chosen a, the initial conditions would have been c' (B2) is true.

5 | FURTHER DISCUSSION

I've shown that the causal model semantics ($>$) described in §2, together with the theory of **Causal Influence** from §3, satisfies (B1) and (B2). Since it clearly satisfies (B3), it must violate the schematic principle (B4)

(B4) If $P > Q$, $P > R$, and $Q \wedge R$ metaphysically necessitates S, then $P > S$.

However, the foregoing does not make it clear *why* the semantics violates this principle.

In this section, I will explain that condition (E1) from **Causal Influence** leads to (B4) being violated. This discussion will put us in a position to appreciate that the theory from §§2–3 will not tell us whether causal counterfactuals like (G1) and (G2) are true or false.

(G1) If you hadn't chosen a, it would have been the case that either the initial conditions were different or there was a miracle.
(G2) If you hadn't chosen a, it would have been the case that the initial conditions were the same and there was no miracle.

5.1 | Principle (B4) and mereological distinctness

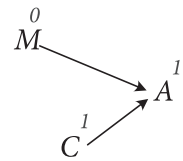
It's tempting to think that the the principle (B4) fails on this semantics because causal influence need not be preserved through metaphysical necessitation. Consider this case: whether the doctor gives morphine causally influences whether the patient dies painlessly. So we get the causal counterfactual 'had the doctor given the patient morphine, they would have died painlessly'. If the patient dies painlessly, this metaphysically necessitates that the patient dies. But whether the doctor gives morphine does not causally influence whether the patient dies. So—you may think—we don't get the causal counterfactual 'had the doctor given the patient morphine, they would have died.'

This thought is tempting but wrong. The reason it is wrong is that the semantics ($>$) does not *require* there to be causal influence between the antecedent and the consequent. Indeed, the counterfactuals (B1) and (B2) are true precisely because there is *not* any causal influence between

whether you choose a and whether there's a miracle, nor between whether you choose A and whether the initial conditions are c. Suppose that, if the doctor fails to give morphine to a dying patient, or gives morphine to a patient who isn't dying, then they will be disciplined. Then, we will have a causal model according to which whether the doctor gives morphine and whether the patient dies both causally influence whether the doctor is disciplined, and whether the doctor gives morphine does not causally influence whether the patient dies. And this causal model will tell us that 'had the doctor given morphine, the patient would still have died' is true. (Or, more trivially, just take a causal model in which both *whether the doctor gives morphine* and *whether the patient dies* are exogenous variables, and there are no endogenous variables or structural equations. This model will tell us that, had the doctor not given morphine, the patient would still have died. And all it will take for this model to be correct is for the exogenous variables to be distinct, which they are, and for them to not causally influence each other, which they do not.)

Instead, the reason that (B4) fails is the requirement of mereological distinctness, (E1). Though there is no causal influence leading from A to M, and no causal influence leading from A to C, we cannot have a causal model which includes *all three* variables, A, M, and C. That is, no system of structural equations like this is correct:

$$M = 0$$
$$C = 1$$
$$A := \overline{M} \cdot C$$



Any system of structural equations which includes all three variables A, M, and C will be incorrect. For the value of A is not fixable independently of the values of M and C. If $M = 0$ and $C = 1$, this metaphysically necessitates that $A = 1$, and it will be impossible for us to set $A = 0$ while M is set to 0 and C is set to 1. That is: if the initial conditions are c and there's no violation of the actual laws of nature, then it will be impossible for you to not choose a. So the variables in $\{A, M, C\}$ are not mereologically distinct, and condition (E1) tells us that no causal model containing these variables can be correct.

Recall, we included the requirement that variables be mereologically distinct to avoid saying that, e.g., had the die not landed high, it would still have landed odd. But ruling out overlapping variables also means that (>) will not say whether the counterfactuals (G1) and (G2) are true or false.

(G1) If you hadn't chosen a, it would have been the case that either the initial conditions were different or there was a miracle.
(G2) If you hadn't chosen a, it would have been the case that the initial conditions were the same and there was no miracle.

According to (>), truth requires a validating model, and falsehood requires a model which validates the negation. When it comes to counterfactuals like (G1) and (G2), there is no causal model which contains variables for both the antecedents and the consequents. So, when it comes to counterfactuals like these, the semantics (>) simply falls silent. It does not say that they are true, nor does it say that they are false. Evaluating counterfactuals like these would require extending the causal modelling semantics (>). It's not clear to me how—or whether—this should be done.

5.2 | Betting on the past and a miracle

Turning to causal decision theory, the discussion from §5.1 above is relevant to decisions like the following.

Betting on the past and a miracle You are certain that the laws are deterministic, that there are not and never will be any miracles, and that the initial conditions were *c*. You are offered a choice between two bets, *a* and *b*. Both bet *a* and bet *b* are bets on the following proposition: there are no miracles, and the initial conditions are *c*. If you take bet *a* and this proposition is true, then you'll win \$10; whereas if you take bet *a* and the proposition is false, you'll get nothing. If you take bet *b* and this proposition is true, then you'll gain \$1; whereas, if you take bet *b* and the proposition is false, then you'll gain \$11.

	There are no miracles and the initial conditions are *c*	**Either there are miracles or the initial conditions are not *c***
Bet *a*	\$10	\$0
Bet *b*	\$1	\$11

Insofar as we think that (G1) is true, we should understand CDT as advising you to take bet *b*. Insofar as we think (G2) is true, we should understand CDT as advising you to take bet *a*. Insofar as we think it's indeterminate which is true, we should understand CDT as implying that it's indeterminate what you should do. Formulating a theory of causal counterfactuals which can tell us whether (G1) and (G2) are true or false would be a complicated task, requiring us to revise some of the formal tools I've been taking for granted here.[38] I won't have much to say about this question, except to register my views that, firstly, it's not pre-theoretically clear what you should do in **Betting on the past and a miracle**; and, secondly, decisions like these aren't potential counterexamples to CDT.

If it is outside of your control whether there are no miracles and the initial conditions are *c*, then it seems clear to me that you should take bet *a*, and that taking bet *b* is irrational, given your beliefs. However, it is not clear to me that this *is* outside of your control. I am inclined to accept a causal counterfactual analysis of what is outside of your control, saying that whether *P* is outside of your control iff *P*'s truth-value does not (causally) counterfactually depend upon how you act. That is: whether *P* is under your control iff there's a choice you could have made such that *P* would have been true, had you made that choice, and there's another choice you could have made such that *P* would have been false, had you made that choice. And whether *P* is outside of your control iff whether *P* is not under your control. While it seems that it is outside of your control whether the initial conditions are *c* or not, and while it seems that it is outside of your control whether the laws are violated or not, it is unclear to me whether the *conjunction* 'the initial conditions are *c* and the deterministic laws are unviolated' is outside of your control.[39]

[38] Woodward (2015) has some helpful suggestions for how this is to be done, but I won't have the space to engage with them here.

[39] When I say that a proposition, *P*, is outside of your control, I just mean that whether *P* is outside of your control.

There's some inclination to think that this *follows* from the initial conditions being outside of your control and the laws being outside of your control. That is, there's some inclination to accept the following schematic principle:

Agglomeration If whether P is outside of your control, and whether Q is outside of your control, then whether $P \wedge Q$ is outside of your control.

Agglomeration is a key premise in van Inwagen (1983)'s consequence argument for incompatibilism—indeed, the same instance of the principle that we are interested in here is the one used in that argument.[40] However, despite its plausibility, **Agglomeration** is false. Consider the following counterexample, from McKay & Johnson (1996): there is a coin which you do not actually flip, but which you could have flipped. Let '$\neg H$' be 'the coin does not land heads', and let '$\neg T$' be 'the coin does not land tails'. Whether the coin lands heads or not is not under your control. There is no choice you could have made such that '$\neg H$' would have been false, had you made that choice. Likewise, whether the coin lands tails or not is not under your control. There is no choice you could have made such that '$\neg T$' would have been false, had you made that choice. But the *conjunction* '$\neg H \wedge \neg T$' *is* under your control. If you were to not flip the coin, this proposition would be true; and if you were to flip it, the proposition would be false (since the coin would either land heads or tails).[41] So **Agglomeration** is false, despite how appealing the principle appears when considered in the abstract.

Suppose that it is not you, but someone else, who faces this decision. And suppose that you know all of the relevant facts, the laws are deterministic, there are no miracles, and the initial conditions are c. You watch this person take bet a and gain \$10. Ask yourself: did this person make the choice which was *objectively* best? That is: did their choice maximise *objective* instrumental value?[42] From my perspective, it is unclear. I'm somewhat tempted to say "taking bet a instead of bet b gained them \$9, since, if they'd taken bet b, they'd have gotten \$1." At the same time, I recognise that there is no way for them to take bet b while the initial conditions remain c and the laws remain unviolated. So there's some inclination to say that taking bet a instead of bet b lost them \$1, since, if they'd taken bet b, either the initial conditions would have been different, or else the laws would have been violated, and so they'd have gotten \$11. It's undeniable that this English counterfactual has a true reading, but I must remind myself that not every English counterfactual is a *causal* counterfactual which reveals genuine control. So I have two conflicting inclinations, neither of which strikes me as dispositive. At the end of the day, I'm just not sure what to say about whether this person's choice has maximised objective instrumental value or not. (Note that all the same considerations hold if the person chooses bet b instead of bet a. In that case, there's some inclination to say that this gained them \$1; and some inclination to say that this lost them \$9. The inclinations are conflicting and inconclusive, and I'm left unsure whether this choice was objectively best.)

According to the causal decision theorist, rational choice is an attempt to maximise objective instrumental value. (You make the attempt by maximising your subjective *expectation* of objective

[40] This is the principle used in what Huemer (2000) calls 'the second version' of the consequence argument. Huemer calls it 'β^*'. As Huemer shows, this alternative formulation is equivalent to the original, in the sense that the principles α^* and β^* used in the second version are equivalent to the principles α and β used in the original argument.

[41] Similar counterexamples are discussed in Widerker (1987) and Huemer (2000).

[42] I take it for granted that there is such a thing as objective instrumental value, though some evidentialists will disagree—see Ahmed & Spencer (2020) and Gallow (ms).

instrumental value.) But in a decision like **Betting on the past and a miracle**, it is unclear which act maximises objective instrumental value in each possible state of the world. So, from the perspective of the causal decision theorist, it is unclear which choice has the highest *expected* instrumental value.

Decisions like these are fascinating. But I think they should not be seen as potential counterexamples to CDT—for at least three reasons. Firstly, because it is not clear what CDT *says* about these decisions. For it is unclear which causal counterfactuals are true in these decisions. Secondly, it is not clear what a decision theory *should* say about decisions like these. For it is unclear which choice maximises objective instrumental value in this decision. Thirdly, it is unclear which choice is objectively best *precisely because* it is unclear which causal counterfactuals are true. And no matter which choice we say is objectively best, CDT will advise you to choose it. Insofar as we have reason to accept the causal counterfactual (G1), we have reason to think that a has less instrumental value than b. In that case, CDT would say that you should take b, which is the choice which is objectively best. And insofar as we have reason to accept the causal counterfactual (G2), we have reason to think that a has more instrumental value than b. In that case, CDT would say that you should take a, which is the choice which is objectively best. So we have a puzzle for our theory of causal counterfactuals; but it is not a challenge to CDT, since, however we resolve the puzzle, CDT will say that you should do what's objectively best.

ACKNOWLEDGEMENTS

Thanks to Melissa Fusco, Jeremy Goodman, Ben Holguín, Matt Mandelkern, and an anonymous referee for helpful conversations and feedback on this material.

Open access publishing facilitated by Australian Catholic University, as part of the Wiley - Australian Catholic University agreement via the Council of Australian University Librarians.

ORCID

J. Dmitri Gallow https://orcid.org/0000-0002-6513-2050

REFERENCES

Ahmed, A. (2013). Causal Decision Theory: A Counterexample. *The Philosophical Review*, 122(2):289–306.

Ahmed, A. (2014). Causal Decision Theory and the Fixity of the Past. *The British Journal for the Philosophy of Science*, 65(4):665–685.

Ahmed, A. (2014). *Evidence, Decision and Causality*. Cambridge University Press, Cambridge, UK.

Ahmed, A., and Spencer, J. (2020). Objective Value is Always Newcombizable. *Mind*, 129(516):1157–1192.

Albert, D. Z. (2003). *Time and Chance*. Harvard University Press, Cambridge, MA.

Beebee, H. (2003). Local miracle compatibilism. *Noûs*, 37(2):258–277.

Bennett, J. (1974). Counterfactuals and Possible Worlds. *Canadian Journal of Philosophy*, 4(2):381–402.

Bennett, J. (1984). Counterfactuals and Temporal Direction. *The Philosophical Review*, 93(1):57–91.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*, Clarendon Press, Oxford.

Braddon-Mitchell, D (2001). Lossy Laws. *Noûs*, 35(2):260–277.

Briggs, R. A. (2012). Interventionist Counterfactuals. *Philosophical Studies*, 160(1):139–166.

DeRose, K. (1999). Can It Be That It Would Have Been Even Though It Might Not Have Been? *Philosophical Perspectives*, 13:385–413.

Dorr, C. (2016). Against counterfactual miracles. *The Philosophical Review*, 125(2):241–286.

Edgington, D. (2004). Counterfactuals and the benefit of hindsight. In P. Dowe, and P. Noordhof editors, *Cause and Chance: Causation in an Indeterministic World*, chapter 2, pages 12–27. Routledge, London.

Elga, A. (2001). Statistical Mechanics and the Asymmetry of Counterfactual Dependence. *Philosophy of Science*, 68S:313–24.

Elga, A. (2022). Confessions of a causal decision theorist. *Analysis*, 82(2):203–213.

Fine, K. (1975). Critical Notice of Lewis, Counterfactuals. *Mind*, 84(335):451–458.
Galles, D., and Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3(1):151–182.
Gallow, J. D. (2016). A Theory of Structural Determination. *Philosophical Studies*, 173(1):159–186.
Gallow, J. D. (ms). Comparativism About Instrumental Value.
Gibbard, A., and Harper, W. L. (1978). Counterfactuals and Two Kinds of Expected Utility. In Hooker, A., Leach, J., and McClennan, E., editors, *Foundations and Applications of Decision Theory*, pages 125–162. D. Reidel, Dordrecht.
Gibbs, C. (2020). Counterfactuals and laws with violations. *Synthese*, 198(11):10643–10659.
Goggans, P. (1992). Do the Closest Counterfactual Worlds Contain Miracles? *Pacific Philosophical Quarterly*, 73(2):137–149.
Goodman, J. (2015). Knowledge, Counterfactuals, and Determinism. *Philosophical Studies*, 172(9):2275–2278.
Hall, N. (2007). Structural Equations and Causation. *Philosophical Studies*, 132(1):109–136.
Halpin, J. F. (1991). The miraculous conception of counterfactuals. *Philosophical Studies*, 63(3):271–290.
Hausman, D. M., and Woodward, J. (1999). Independence, Invariance, and the Causal Markov Condition. *The British Journal for the Philosophy of Science*, 50(4):521–583.
Hiddleston, E. (2005). A Causal Theory of Counterfactuals. *Noûs*, 39(4):632–657.
Hoffmann-Kolss, V. (2021). Interventionism and Non-Causal Dependence Relations: New Work for a Theory of Supervenience. Australasian Journal of Philosophy.
Huber, F. (2013). Structural Equations and Beyond. *The Review of Symbolic Logic*, 6(4):709–732.
Huemer, M. (2000). Van Inwagen's Consequence Argument. *The Philosophical Review*, 109(4):525–544.
Jackson, F. (1977). A causal theory of counterfactuals. *Australasian Journal of Philosophy*, 55(1):3–21.
Joyce, J. M. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press, Cambridge.
Joyce, J. M. (2016). Arif ahmed: Evidence, decision and causality. *The Journal of Philosophy*, 113(4):224–232.
Kment, B. (2006). Counterfactuals and Explanation. *Mind*, 115(458):261–309.
Kment, B. (ms). Decision, Causality, and Pre-Determination.
Kutach, D. (2002). The entropy theory of counterfactuals. *Philosophy of Science*, 69(1):82–104.
Lange, M. (2000). *Natural Laws in Scientific Practice*. Oxford University Press, Oxford.
Lewis, D. K. (1973). *Counterfactuals*. Blackwell Publishers, Malden, MA.
Lewis, D. K. (1979). Counterfactual Dependence and Time's Arrow. *Noûs*, 13(4):455–476.
Lewis, D. K. (1980). A Subjectivist's Guide to Objective Chance. In Jeffrey, R. C., editors, *Studies in Inductive Logic and Probability*, volume II, pages 263–293. University of California Press, Berkeley.
Lewis, D. K. (1986). Events. In *Philosophical Papers*, volume II, pages 241–269. Oxford University Press, New York.
Lewis, D. K. (1986). Postscripts to 'causation'. In *Philosophical Papers*, volume II. Oxford University Press, New York.
Lewis, D. K. (2020). To Jonathan Bennett, 21 April 1981. In Beebee, H. and Fisher, A., editors, *Philosophical Letters of David K. Lewis*. Oxford University Press, Oxford.
Loewer, B. (2007). Counterfactuals and the Second Law. In Price, H. and Corry, R., editors, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, chapter 11, pages 293–326. Oxford University Press, Oxford.
Maudlin, T. (2007). A Modest Proposal Concerning Laws, Counterfactuals, and Explanations. In *The Metaphysics within Physics*, pages 5–49. Oxford University Press, Oxford.
McKay, T. J., and Johnson, D. (1996). A Reconsideration of an Argument against Compatibilism. *Philosophical Topics*, 24(2):113–122.
Nute, D. (1980). *Topics in Conditional Logic*. D. Reidel, Dordrecht.
Paul, L. A., and Hall, N. (2013). *Causation: A User's Guide*. Oxford University Press, Oxford.
Sandgren, A., and Williamson, T. L. (2021). Determinism, Counterfactuals, and Decision. *Australasian Journal of Philosophy*, 99(2):286–302.
Schaffer, J. (2004). Counterfactuals, Causal Independence and Conceptual Circularity. *Analysis*, 64(4):299–309.
Slote, M. A. (1978). Time in Counterfactuals. *The Philosophical Review*, 87(1):3–27.
Sobel, J. H. (1994). *Taking Chances: Essays on Rational Choice*. Cambridge University Press, Cambridge.
Solomon, T. C. P. (2021). Causal decision theory's predetermination problem. *Synthese*, 198(6):5623–5654.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. The MIT Press, Cambridge, MA, 2nd edition.

Stalnaker, R. C. (1968). A Theory of Conditionals. In Rescher, N., editors, *Studies in Logical Theory*, chapter 4, pages 98–112. Oxford University Press, Oxford.

Stalnaker, R. C. (1980). A Defense of Conditional Excluded Middle. In Harper, W. L., Stalnaker, R., and Pearce, G., editors, *Ifs*, pages 87–104. D. Reidel, Dordrecht.

Stalnaker, R. C. (1981). Letter to David Lewis. In Harper, W., Stalnaker, R., and Pearce, G., editors, *Ifs*, pages 151–152. D. Reidel Publishing Company, Dordrecht.

van Inwagen, P. (1983). *An Essay on Free Will*. Oxford University Press, New York, NY.

Walters, L., and Williams, J. R. G. (2013). An argument for conjunction conditionalization. *The Review of Symbolic Logic*, 6(4):573–588.

Wasserman, R. (2006). The Future Similarity Objection Revisited. *Synthese*, 150(1):57–67.

Widerker, D. (1987). On an Argument for Incompatibilism. *Analysis*, 47(1):37–41.

Williamson, T. L., and Sandgren, A. (forthcoming). Law-Abiding Causal Decision Theory. The British Journal for the Philosophy of Science.

Wilson, J. M. (2014). Hume's Dictum and the asymmetry of counterfactual dependence. In Wilson, A., editors, *Chance and Temporal Asymmetry*, pages 258–279. Oxford University Press, Oxford.

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

Woodward, J. (2015). Interventionism and causal exclusion. *Philosophy and Phenomenological Research*, 91(2):303–347.

How to cite this article: Gallow, J.D. Causal counterfactuals without miracles or backtracking. *Philosophy and Phenomenological Research*. 2022;1–31. https://doi.org/10.1111/phpr.12925