A counterfactual explanation for recency effects in double prevention scenarios: commentary on Thanawala and Erb (2024)

Tadeg Quillien[1], Kevin O'Neill[2], and Paul Henne[3]
[1]Department of Psychology, University of Edinburgh
[2]Department of Psychology & Neuroscience, Center for Cognitive Neuroscience, Duke University
[3]Department of Philosophy, Neuroscience Program, Lake Forest College

Many cognitive scientists and philosophers take cases of double prevention to be one of the primary motivations for accepting causal pluralism, the view that people have multiple concepts of causation. Thanawala and Erb (2024) argue against Lombrozo's (2010) account of causal pluralism. They find that the temporal order of events affects people's causal judgments in double prevention cases, and they argue that this finding is not easily explained by prominent versions of causal pluralism or by counterfactual theories. In contrast to this interpretation, we argue that counterfactual thinking can explain their findings. On this explanation, the temporal order of events affects the extent to which people simulate counterfactual alternatives to these events. We show that under this assumption, a recent counterfactual model of causal judgment can reproduce all qualitative effects of temporal order found in Thanawala and Erb's (2024) new work. Our findings complement past research that applied the counterfactual framework to temporal-order effects and double prevention cases independently, suggesting that these explanations are highly generalizable.

Keywords: causal judgment, counterfactual thinking, recency, double prevention, computational modeling, modal cognition

Pam threw a ball at a window. Tom was just about to block the ball from hitting the window when John accidentally fell into Tom, preventing Tom from preventing the ball from hitting the window. So, the ball hit the window, and it shattered.

This is a case of double prevention. A productive cause like Pam throwing the ball at the window initiates a sequence of events. A possible preventer like Tom nearly preventing the ball from hitting the window possibly prevents the productive cause from bringing about the outcome. And a double preventer like John falling into Tom prevents the possible preventer from preventing the outcome. Then, an outcome, like the window shattering, happens. People tend to judge that the productive cause, like Pam throwing the ball, caused the window to shatter, and they tend to deny that the double preventer, like John falling into Tom, caused the window to shatter (Henne & O'Neill, 2022; Lombrozo, 2010).

Cases of double prevention are famously inconsistent with one of the best models of causal judgment, counterfactual models (Hall, 2002, 2004). On counterfactual models, when people are making a judgment about the extent to which an event caused an outcome, they think about how things could have been different. So, when determining whether an event is a cause, they imagine that the potential cause did not happen and ask if the outcome would have happened in that imagined scenario. If the outcome would not have happened in that scenario, then that potential cause made a difference—it caused the outcome. For example, suppose that someone wants to know whether Pam throwing the ball caused the window to shatter. On a counterfactual view, this person would imagine that Pam did not throw the ball at the window. In this imagined scenario, the window would not have shattered. So, Pam throwing the ball at the window caused the window to shatter.

To see why cases of double prevention are inconsistent with counterfactual theories, consider again the case above. On counterfactual accounts, the productive cause—like Pam throwing the ball at the window—is a cause of the window shattering, as we just explained. But, on these kinds of accounts, the double preventer is also a cause of the outcome; in these cases, if the double preventer did not happen, the outcome would not happen, too. For instance, if John had not fallen into Tom, Tom would have prevented the ball from hitting the window, and the window would not have shattered. So, a standard counterfactual account predicts that both the productive cause and the double preventer caused the outcome. But this is inconsistent with people's causal judgments: as we mentioned above, people tend to judge that the productive cause and not the double preventer caused the outcome in double-prevention cases (Henne & O'Neill, 2022; Lombrozo, 2010). In other words, counterfactual accounts seem insufficient to explain the pattern of people's causal judgments in cases of double prevention.

As such, many philosophers and cognitive scientists take this pattern of judgments to support causal pluralism, the idea that there is both a productive and a counterfactual concept of causation (for discussion, see Hall, 2004; Henne, 2023; Lombrozo, 2010). On this kind of pluralist view, a productive concept of causation explains why people tend to judge the productive cause as more causal than the double preventer; the productive cause transmits energy to the outcome, while the double preventer does not at all. Some argue then that the extent to which people judge that the double preventer is a cause of the outcome depends on the weight or salience of the counterfactual alternative (Lombrozo, 2010).

In an important new paper, Thanawala and Erb (2024) challenge Lombrozo's (2010) pluralist account. Thanawala and Erb (2024) find that the order of events in double-prevention scenarios affects people's causal judgments (see also, Experiment 3 in Henne & O'Neill, 2022). To see this, consider a variation of our case from before. Suppose Tom was already blocking the ball from hitting the window, and then Pam threw the ball at the window. John then accidentally fell into Tom, preventing Tom from preventing the ball from hitting the window. So, the ball hit the window, and it shattered. In cases ordered like this one, Thanawala and Erb (2024) found that people are more inclined to judge that the double preventer caused the outcome.

Thanawala and Erb 2024 argue that this finding (and others) challenge Lombrozo's pluralist account because her account predicts that people should not tend to judge that unintentional double preventers caused the outcome. They also argue that the temporal-order effects in these cases give us reason to think that process information play an important role in causal judgment (Thanawala & Erb, 2024).

We suggest that these effects of temporal order can be naturally explained within a counterfactual framework. According to counterfactual theories, causal judgments are influenced by the counterfactual possibilities that people consider (Gerstenberg et al., 2017, 2021; Icard et al., 2017; Krasich et al., 2024; Quillien, 2020). And people are more likely to consider counterfactual alternatives to some events rather than others (Byrne, 2016; Kahneman & Miller, 1986). Some factors, moreover, tend to affect people's tendency to consider certain counterfactuals (Byrne, 2016). In particular, people tend to consider counterfactual alternatives to more recent events relative to those that happened earlier in time (Byrne, 2016; Byrne et al., 2000; Henne et al., 2021; Segura et al., 2002; Walsh & Byrne, 2004).

With these assumptions, counterfactual models can explain temporal ordering effects in causal judgment. In particular, Henne and colleagues (2021) showed that a recency effect in counterfactual reasoning can parsimoniously explain why causal judgment sometimes exhibits a recency effect and sometimes a primacy effect and in what circumstances they do so. Generally, it seems that counterfactual accounts can explain effects of temporal order on causal judgment (see also Henne, 2023; Ziano and Pandelaere, 2022; for background on temporal effects in causal judgment see Hart and Honoré, 1985; Hilton et al., 2010; McClure et al., 2007; Spellman, 1997).

Recent work also suggests that double prevention cases might actually be consistent with a counterfactual framework, despite the arguments we reviewed earlier (Henne & O'Neill, 2022; O'Neill et al., 2022). According to recent computational models, people make causal judgments by i) considering several counterfactual alternatives to what happened and then ii) assessing how much the outcome depends on a given cause across these counterfactual possibilities (Icard et al., 2017; Quillien, 2020). These models can explain people's causal judgments in double-prevention cases (Henne & O'Neill, 2022; O'Neill et al., 2022). Notably, they can explain why people often view the productive cause as more causal than the double preventer. Consider, for example, the productive cause like Pam throwing the ball at the window. The productive cause is always necessary for the outcome; for instance, if Pam had not thrown the ball at the window, the window would not have shattered. The double preventer like John falling into Tom, however, is not necessary for the outcome in the same sense; in an alternative possibility where there is no possible prevention, the outcome can happen even in the absence of the double preventer. For example, if John had not fallen into Tom in a situation where Tom was never able to block the ball that Pam threw, the window could have still shattered.

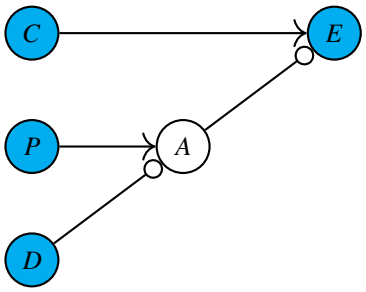
With this background in mind, the new patterns of causal judgment in double-prevention cases reported by Thanawala and Erb (2024) might be consistent with a counterfactual account of human causal judgment. In order to test our hypothesis formally, we explore the predictions of a computational model of causal judgment, the Counterfactual Effect Size model (Quillien, 2020; Quillien & Lucas, 2023), in the context of Experiments 3 and 4 in Thanawala and Erb (2024). Specifically, we hypothesized that (a) temporal order does not directly influence causal judgments, rather such effects are mediated through counterfactual simulation, and that (b) in particular, people are more inclined to imagine counterfactuals to recent events rather than earlier events. By comparing the Counterfactual Effect Size model to modified versions of the model, we provide independent support for both of these hypotheses.

Model

Thanawala and Erb (2024) asked participants to read vignettes about the following story. Anna puts a package on a conveyor belt that leads to a shipping dock. Further down on the conveyor belt there is a scanner that would block the package from reaching its destination because the package does not have the required label. Brad turns on the scanner.

Clara disconnects the scanner from its power outlet, preventing it from preventing the package from reaching its destination. The package then reaches its destination.

We refer to Anna's action as the productive cause (C), Brad's action as the possible preventer (P) and Clara's action as the double preventer (D). We also use the variable A to refer to the actualization of the possible preventer—that is, the counterfactual event where Brad successfully blocks the package. Finally, the outcome variable E represents the package successfully reaching its destination. Although Thanawala and Erb (2024) asked participants about the causal role of *agents*, we assume that A, P, and D represent *events*. On this assumption, D is the event of Clara disconnecting the scanner.[1] The causal structure is illustrated in the graph below where events that actually happened are in blue and where lines with nodes at the end represent prevention relationships:



In their experiments, Thanawala and Erb (2024) manipulated the temporal order of events in the double-prevention case above. For example, in the PCD condition, the possible preventer, P, happens first (Brad turns on the scanner); then the productive cause, C, happens (Anna puts the package on the conveyor belt); then the double preventer, D, happens (Clara disconnects the scanner); and then the outcome, E, happens (the package successfully reaches its destination). Thanawala and Erb (2024) tested the following orderings: PCD, PDC, CPD, and CDP.

Overview of the model

We will explore the predictions of a counterfactual computational model of causal judgment called the Counterfactual Effect Size (CES) model. The CES model assumes that people make causal judgments by simulating counterfactual alternatives to what happened and then computing a measure of effect size that quantifies how much changing the value of a potential cause changes the value of the outcome across the imagined counterfactual alternatives (Quillien, 2020; Quillien & Lucas, 2023). On the CES model, the double preventer caused the outcome to the extent that, across the counterfactual alternatives that people imagine, there is a high correlation between the double preventer happening and the outcome happening. For instance, Clara disconnecting the scanner caused the package to reach its destination to the extent that there is a high correlation between Clara disconnecting the scanner and the package reaching its destination across the counterfactual alternatives that people imagine.

Effect size measure

We consider a distribution $\mathcal{S}$ over counterfactuals worlds, which can be constructed by sampling possible worlds from the following generative process (see Henne & O'Neill, 2022):

$$
\begin{aligned}
C &\sim \mathcal{S}(C) \\
P &\sim \mathcal{S}(P) \\
D &\sim \mathcal{S}(D) \\
A &:= P \ \wedge \ \neg D \\
E &:= C \ \wedge \ \neg A
\end{aligned}
$$

In words, the first three lines state that we stochastically sample C, P, and D (see below), and the last two lines are structural equations that determine the values of A and E as a function of the value of their causes (Pearl, 2009).

From this distribution $\mathcal{S}$ over counterfactual worlds, we can compute a measure of causal strength quantifying the extent to which an event X caused the outcome E. In a double-prevention case, the causal strength $\kappa_{X \to E}$ of event X is equivalent to the correlation between X and the outcome E, across counterfactual worlds (O'Neill et al., 2022; Quillien, 2020).

Determinants of counterfactual sampling

Variables C, P and D are *exogenous* in the sense that they do not depend on the values of other variables. In the CES model, an exogenous variable X is sampled according to its *sampling propensity* $\mathcal{S}(X)$:

$$X \sim \mathcal{S}(X)$$

where $\mathcal{S}(X)$ depends on the actual-world value of X and its normality, $Pr(X)$.[2] That is, each time we sample X we either keep the value it has in the actual world with probability s_t, or we re-sample it from a prior distribution $Pr(X)$. Formally:

$$\mathcal{S}(X = x) = s_t \delta(x) + (1 - s_t) Pr(X = x)$$

[1] This is a standard assumption in the causal modeling literature (see, for instance, Icard et al., 2017).

[2] The normality of an event is a function of its statistical probability, but it can also be sensitive to other factors such as its conformity to prescriptive or functional norms (see Icard et al., 2017; Kominsky & Phillips, 2019). Here, we do not model these factors explicitly. Instead, we infer normality parameters from participants' causal judgments.

where $\delta(x) = 1$ if $X = x$ in the actual world and 0 otherwise (Lucas & Kemp, 2015; Quillien & Lucas, 2023; Quillien et al., 2023).

To implement our assumption that temporal ordering might influence counterfactual simulation, we let the stability parameters s_t depend on the temporal order of the event. Specifically, we assume that the first, second, and third events in a scenario have stability parameter s_1, s_2, and s_3, respectively. Stability parameters determine the probability that the model will leave a variable at the value it has in the actual world rather than re-sampling a counterfactual value. So, the greater the stability, the more likely the value of a variable will remain unchanged. Our hypothesis is that people tend to imagine events that happened earlier in time happening just as they did and that people tend to imagine more recent events happening differently.

We treat $Pr(X)$ and s_t as free parameters. By hypothesis, participants should be more likely to re-sample recent events such that $s_1 > s_2 > s_3$. Inferring the stability parameters s_t from the data allows us to test this hypothesis. We model $Pr(X)$ as a free parameter because it was not manipulated or measured in Thanawala & Erb's experiment. Importantly, we set the values of all parameters to be fixed across conditions. For example, the value of $Pr(D)$ is the same in all conditions–regardless of whether D was the first, second, or third event–and the value of s_2 is the same whether the event is C, P, or D. In this way, the model cannot arbitrarily account for differences across conditions by adjusting parameters.

Methods

We fit the CES model to the combined data from Experiment 3 and 4 in Thanawala and Erb (2024) using the `cmdstanr` interface to the probabilistic programming language Stan (Gabry et al., 2024; Stan Development Team, 2024). We sampled 10,000 iterations across four chains of Hamiltonian Monte-Carlo. We used an ordered probit likelihood to map model-produced judgments to the ordinal scale of the data.

We chose uniform priors over the prior probability of each event ($Pr(X)$), standard normal priors on the logit-scaled stability parameters of each event (s_t), and standard normal priors on the ordered probit intercepts. For all parameters, we report posterior medians and 95% credible intervals.

We fit three models to the data. The *full* CES model included a time-dependent stability parameter as we described above. In addition to this model, we fit two alternative models. The purpose of these alternative models is to determine whether simpler assumptions than our full model can explain these data. These models are also built with the CES model at their core, but they make different assumptions about the effect of temporal order.

First, we fit a *primacy* model in which the stability parameters were not allowed to decrease with recency such that $s_1 \leq s_2 \leq s_3$. In other words, the primacy model assumes that people are more likely to simulate counterfactual alternatives to *earlier* events rather than more recent events (or that temporal order does not affect counterfactual simulation). We include this model because it provides an alternative to our hypothesis that people prefer to imagine alternatives to *recent* events. If the primacy model accounts for the data less well than the full model, this would suggest that a preference for simulating alternatives to recent events is necessary to explain the pattern of people's causal judgments in these cases.

Second, we fit a *bias* model in which temporal recency exerted a direct effect on causal judgments rather than an indirect effect on counterfactual simulation. Specifically, in this model the actual causal strength of each event was adjusted by the temporal order t of the event such that $\kappa^*_{X\to E} = \kappa_{X\to E} + \beta t$, where β is a free parameter. Under this structure, positive values of β would indicate that people have a general preference to identify recent events as causes. To remove the indirect effect of temporal recency on counterfactual simulation, we fixed the stability parameters in this model over time (i.e., $s_1 = s_2 = s_3$). Overall, the bias model assumes that while participants may be more inclined to identify recent events as causes of an outcome, this tendency is due to a response bias instead of an influence on counterfactual simulation. We included this model to test whether counterfactual reasoning is necessary to explain the effect of temporal order. So, if the effect of temporal order on causal judgments is indeed mediated by counterfactual thinking, this model should also make worse predictions than the full CES model.

R code for analysis is available on the Open Science Framework at `https://osf.io/acmsh/?view_only=381b9166ab524596bd07ba15fb0ee976`.

Results

Parameter estimates

As predicted, the best-fitting values of the stability parameters s_t for the full CES model decreased with recency (s_1 = .92, 95% CI = [.83, .97]; s_2 = .71, 95% CI = [.62, .79]; s_3 = .17, 95% CI = [.07, .31]; see Figure 1). This pattern is consistent with the hypothesis that participants were more likely to simulate counterfactual alternatives to recent events relative to earlier events.

In the primacy model, best-fitting values of s_t saw a modest increase with recency (s_1 = .03, 95% CI = [.01, .05], s_2 = .03, 95% CI = [.02, .06], s_3 = .26, 95% CI = [.17, .37]). The bias model estimated that participants had a general tendency to rate recent events as more causal (β = .27, 95% CI = [.22, .32], s = .53, 95% CI = [.18, .85]). Parameter estimates of prior probabilities for each model are listed in Table 1.

Model	Parameter	Posterior Estimate
Full	$Pr(C)$	.87 [.69, .97]
	$Pr(D)$	.68 [.35, .92]
	$Pr(P)$	.09 [.01, .24]
Primacy	$Pr(C)$	.99 [.97, 1.00]
	$Pr(D)$	.003 [.0002, .02]
	$Pr(P)$	.61 [.45, .76]
Bias	$Pr(C)$	.61 [.19, .92]
	$Pr(D)$	.28 [.02, .77]
	$Pr(P)$	.57 [.08, .95]

Table 1

Posterior medians and 95% credible intervals of the prior probability parameters from each model. Actual sampling probabilities were computed as a linear combination of these prior probabilities and the actual-world values of events weighted by stability (see section: Determinants of counterfactual sampling*).*

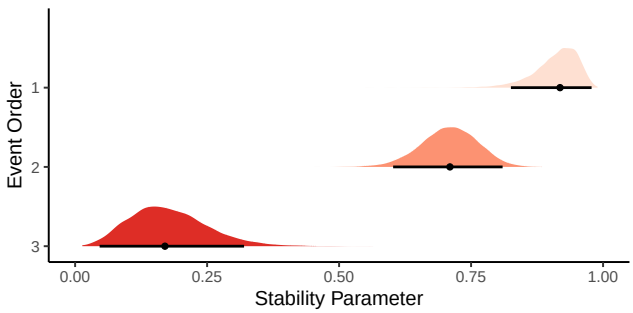


Figure 1

Posterior distributions over the stability parameters from the full CES model. Although they were unconstrained during model fitting, the stability parameters decrease with time, suggesting that people are more inclined to simulate counterfactuals to recent events. Points indicate posterior medians and errorbars indicate 95% credible intervals.

Model fit

We found that the full CES model can account for the effect of temporal order on causal judgment across conditions, while the alternative models cannot (see Figure 2). We assessed model fit in two ways. First, we computed the itemwise Pearson correlation coefficient between model predictions and average participant judgments. Second, we computed the expected log pointwise predictive density using approximate leave-one out cross-validation (*elpd_loo*, see Vehtari et al., 2017). This measure naturally controls for model complexity and, therefore, penalizes models with more free parameters.

The full CES model made predictions that qualitatively matched the data well ($r(6)$ = .88, 95% CI = [.83, .92], *elpd_loo* = -3802.4, *SE* = 32.3). In all four conditions, it estimated higher causal judgments for the factor preferred by participants.

The primacy model provided a less complete description of the data ($r(6)$ = .64, 95% CI = [.56, .71], *elpd_loo* = -3828.2, *SE* = 31.9, Δ*elpd_loo* = -25.9, *SE* = 8.4). In particular, it predicted that participants would prefer the double preventer as the cause of the effect in all four conditions. While this prediction was accurate in the PCD and CPD conditions, it was inaccurate in the PDC and CDP conditions.

Finally, the bias model also provided a less complete description of the data than the full CES model ($r(6)$ = .81, 95% CI = [.79, .81], *elpd_loo* = -3812.4, *SE* = 32.1, Δ*elpd_loo* = -10.0, *SE* = 6.5). While the bias model made accurate predictions in most of the experimental conditions, it incorrectly predicted that participants would prefer the double preventer as the cause of the effect in the CDP condition.

Interpretation

There is an intuitive explanation for why the full model makes the predictions it does. Consider first the PCD and CPD condition, where the full model predicts that the double preventer is the most causal event. In these conditions, the double preventer, D, happens after the productive cause, C. Because the productive cause happens early, the model rarely re-samples it when simulating counterfactuals. By contrast, the model often re-samples whether the double preventer happens because it happens more recently (Figure 3). And when the doubler preventer does not happen, the outcome is prevented from happening. Across these counterfactuals, there is, therefore, a high correlation between whether the double preventer happens and whether the outcome happens that accounts for the higher causal judgments about the double preventer.

To see this clearly, consider the CDP ordering. In the CDP condition, the possible preventer, P happens last, and, therefore, the model has a high tendency to re-sample whether the possible preventer happens (Figure 3). This means that in many counterfactual alternatives, the possible preventer does not happen (for instance, Brad does not try to start the scanner). Therefore, it does not matter whether the double preventer, D, happens or not because the outcome, E, happens frequently in scenarios where it is not prevented by P. For instance, it does not matter whether Clara disconnects the scanner or not because people tend to imagine Brad not trying to start the scanner and the package will just arrive at its destination in those scenarios. So, across counterfactuals, there is a relatively low correlation between the double preventer and the outcome, which accounts for the relatively lower causal judgments about the double preventer in this case. Note that by contrast, the bias model is incapable of

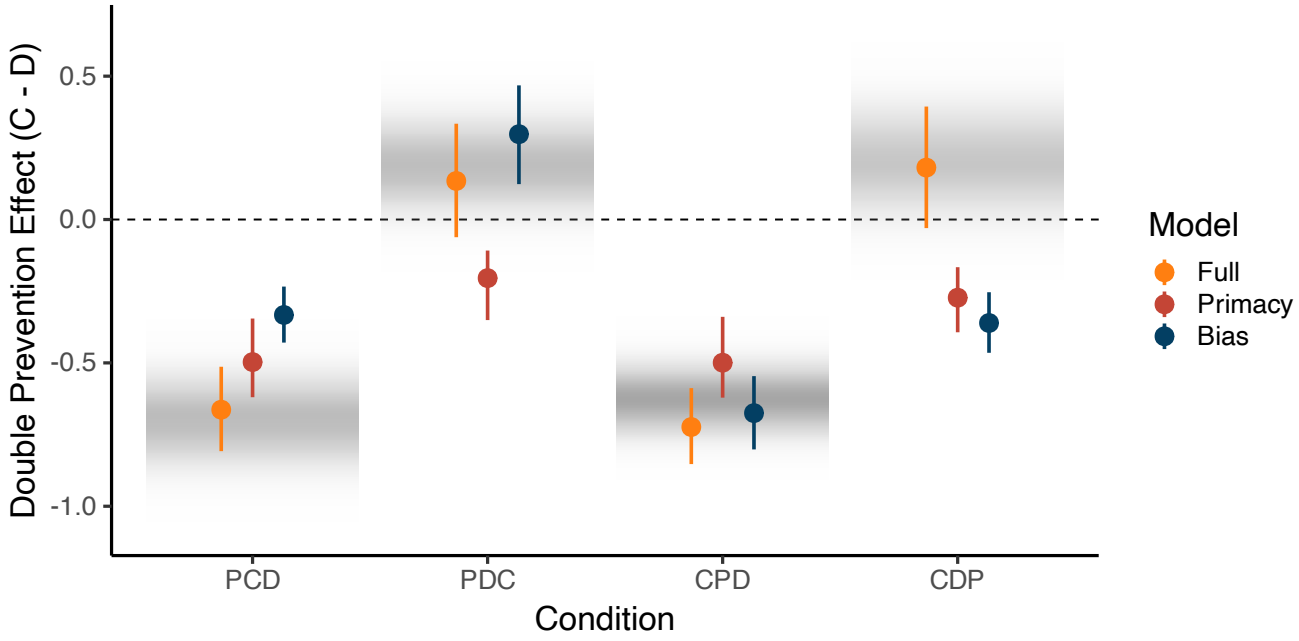


Figure 2

Posterior contrasts of mean causal judgment by factor from each model (color) alongside average human causal judgments (black shading) from Thanawala and Erb (2024). Positive values indicate preference for the productive cause (C), and negative values indicate preference for the double preventer (D). Points indicate posterior medians and errorbars indicate 95% credible intervals. Shaded black regions depict human judgments (contrasts from a fully parameterized ordinal regression model with factor, condition, and their interaction as predictors). Conditions are labeled according to the temporal ordering of events. For instance, in the PCD condition, the possible preventer (P) is first, the productive cause (C) is second, and the double preventer (D) is third.

explaining people's judgments in the CDP condition because it predicts that the double preventer should be more causal than the productive cause. The bias model makes this prediction because it infers that people have a general tendency to view recent events as more causal.

Interestingly, our parameter estimates suggest that people tend to imagine the possible preventer not happening more than they imagine the other two events not happening. This tendency might reflect a bias to think of counterfactual situations where elements of a system function as they are supposed to function (see Hitchcock & Knobe, 2009; Kominsky & Phillips, 2019). In the story used by Thanawala and Erb (2024), the package sent by Anna should ideally reach the shipping dock, given the way the shipping facility is supposed to work. Because preventing that outcome would deviate from the normal functioning of the shipping facility, participants might be inclined to think that normal possibilities are those in which the possible preventer does not happen (e.g., possibilities in which the scanner that would block the package does not turn on). Empirical tests of this explanation are a fruitful direction for future research.

Discussion

Thanawala and Erb (2024) found that the temporal order of events in double prevention cases affect people's causal judgments. These data are relevant to an ongoing debate in causal cognition about whether double prevention scenarios show that people have several concepts of causation or just one (for discussion, see Henne, 2023).

One side of this debate suggests that counterfactual models explain people's causal judgments in double prevention cases without appealing to a plurality of causal concepts (Henne, 2023; Henne & O'Neill, 2022; O'Neill et al., 2022). This work relies on computational models of causal judgment, according to which people consider several counterfactual possibilities when making a causal judgment (Icard et al., 2017; Quillien, 2020; Quillien & Lucas, 2023). Some of these models hold that the outcome covaries more highly with the productive cause than with the double preventer across the counterfactual possibilities that people consider in standard double-prevention cases. This fact explains why people often view the productive cause as more causal than the double preventer (O'Neill et al., 2022).

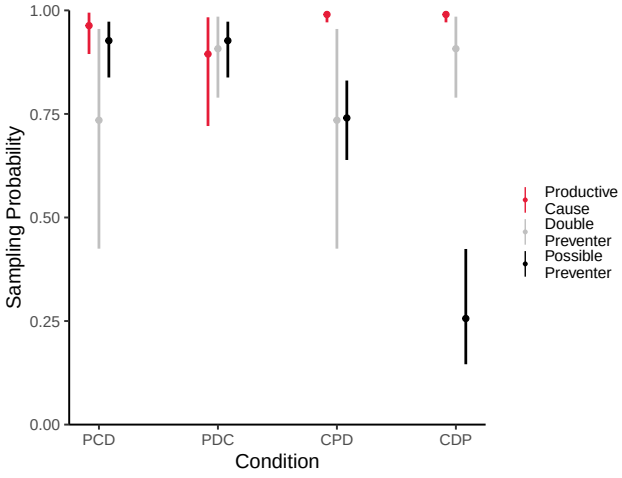


Figure 3

Implied sampling propensities of each event per condition estimated by the full CES model. High values indicate a high probability that the corresponding variable takes value 1 in a simulated counterfactual world. Points indicate posterior medians and errorbars indicate 95% credible intervals. Note that these sampled probabilities were not fitted independently from each other; they were computed as a linear combination of the corresponding normality and stability parameters (see Figure 1 and Table 1).

Here, we find that this counterfactual account can also explain the new effects of temporal order reported by Thanawala and Erb (2024). Analyzing these data using a computational approach, we find that participants appear to preferentially simulate counterfactual alternatives to recent events—a result that parallels existing empirical findings in counterfactual reasoning (Byrne et al., 2000; Henne et al., 2021; Segura et al., 2002; Walsh & Byrne, 2004). Given this recency effect in counterfactual simulation, a recent counterfactual model of causal judgment, the CES model, can reproduce all of the effects of temporal order found by Thanawala and Erb (2024).

These results bolster the counterfactual account of people's causal judgments in double prevention cases (Henne & O'Neill, 2022). One appealing feature of this account is that it provides a unified explanation of people's causal judgments: the same cognitive processes underlie people's judgments about the productive cause and the double preventer (for discussion, see Henne, 2023).

Our proposal also accords with a large body of work on counterfactual models of causal judgment. Counterfactual theories like the CES model are supported by empirical findings outside of the context of temporal order or double prevention (e.g. Konuk et al., 2023; Morris et al., 2019; Quillien & Barlev, 2022; Quillien & Lucas, 2023). In the context of double prevention, Henne and O'Neill (2022) and O'Neill et al. (2022) have shown that recent counterfactual models can predict the contexts in which people judge a double preventer as less or more causal than the productive cause of an outcome. In the domain of temporal order, Henne et al. (2021) found that counterfactual models can explain why people sometimes judge recent events as more causal than early events and sometimes judge early events as more causal. The fact that the theories in this previous work can account for empirical data reported by Thanawala and Erb (2024) suggests that their explanatory power naturally generalizes to novel contexts.

Because we explain people's causal judgments in terms of a unified process of counterfactual reasoning, our results actually bolster Thanawala & Erb's challenge to Lombrozo's (2010) causal pluralist account. We note that Lombrozo (2010) herself suggests an alternative, non-pluralist account of causal judgment in the discussion of her original paper. According to her exportable dependence view, people favor causes that would have led to the outcome even if the background circumstances had been different (Lombrozo, 2010). According to this view, people attribute less causal responsibility to double preventers because double preventers will bring about an outcome less reliably than a productive cause (Lombrozo, 2010). In fact, the CES model was partly inspired by the exportable dependence view, and can be seen as a computational implementation of the idea (Quillien, 2020). From this perspective, our findings accord with some of the ideas put forward in Lombrozo's original paper.

Our account, moreover, is consistent with some of Thanawala and Erb's 2024 general perspective. In their work, they suggest that their acount can be viewed as a process-grounded dependency theory in which "process-based considerations (e.g., the order in which each character acts)" guide the generation of possible worlds (Thanawala & Erb, 2024, p. 13). On our counterfactual account, temporal order guides the generation of the alternative possibilities that people consider such that people are more inclined to consider alternative possibilities to recent events relative to those that happened earlier in time. From this perspective, recent counterfactual models do not eliminate all the insights of causal pluralism; some process information impacts causal judgment, but this role is mediated by counterfactual thinking (Henne, 2023). As such, the counterfactual account we discuss here rejects causal pluralism's claim that there are two different concepts of causation, while allowing for process information to affect causal judgments.

Thanawala and Erb also argue that people's causal judgments in double-prevention cases might depend on whether people mentally represent the possible preventer as capable of posing a threat to the outcome (Thanawala & Erb, 2024, p. 13). In the PDC case, for instance, Clara disconnects the scanner before the package has even been put on the con-

veyor belt (i.e. she disables the possible preventer before the productive cause happens). In such a situation, the threat of the possible preventer seems remote, so people might not consider it a threat at all. And this consideration might explain why people do not ascribe much causal responsibility to the double preventer—the double preventer is ineffective (Thanawala & Erb, 2024).

This view is slightly different from our proposal, but it is conceptually similar. Within our counterfactual framework, this conjecture could be implemented by assuming that participants in the PDC condition, for instance, tend not to imagine the successful prevention of the scanner blocking the package. In imagined scenarios that lack the successful prevention, the outcome would happen whether or not the double prevention happened. So, the double preventer will only be weakly correlated with the outcome across imagined possibilities. On our account then, this tendency not to imagine the successful prevention would explain participants' reluctance to see the double preventer as a cause in such cases.

The main difference between Thanawala and Erb's (2024) account and ours is that they focus on whether people represent a relation between two variables (whether the possible preventer P is a possible threat to the outcome E), while we focus on how people simulate individual variables. Our approach is convenient for modeling how temporal information influences counterfactual simulation (Byrne et al., 2000; Henne et al., 2021; Segura et al., 2002; Walsh & Byrne, 2004), but Thanawala and Erb (2024)'s alternative proposal remains a fruitful direction for future research.

Notably, Thanawala and Erb (2024) also argue against a counterfactual interpretation of their data on the basis of their participants' answers to a comprehension question. In their experiments, they asked participants whether the outcome would still have happened in the absence of the productive cause, and participants overwhelmingly answered (correctly) that it would not have. And participants answered in this way even in conditions where they assigned low causal ratings to the productive cause. Thanawala and Erb (2024) take this finding to be evidence against a counterfactual account.

This finding may be inconsistent with simple counterfactual theories, but it is perfectly consistent with more sophisticated, recent accounts (Icard et al., 2017; Quillien, 2020). On these accounts, an event that is necessary for the outcome, for instance, can still have low causal strength if it is not robustly sufficient for the outcome (i.e., if there are many counterfactual possibilities where the event happens and the outcome does not; Icard et al., 2017; Quillien, 2020). In the PCD condition, for example, our model assigns a low causal rating to the productive cause because it simulates many counterfactual possibilities where the productive cause happens but the outcome does not (because the double preventer is absent). Crucially, the model gives a low causal rating to the productive cause even though it correctly represents the fact that the outcome would not have occurred in its absence. As such, more recent counterfactual models are consistent with this finding.

References

Byrne, R. M. (2016). Counterfactual thought. *Annual review of psychology*, *67*(1), 135–157.

Byrne, R. M., Segura, S., Culhane, R., Tasso, A., & Berrocal, P. (2000). The temporality effect in counterfactual thinking about what might have been. *Memory & Cognition*, *28*, 264–281.

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2024). *Cmdstanr: R interface to 'cmdstan'* [R package version 0.8.1, https://discourse.mc-stan.org]. https://mc-stan.org/cmdstanr/

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological review*, *128*(5), 936.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological science*, *28*(12), 1731–1744.

Hall, N. (2002). Non-locality on the cheap? a new problem for counterfactual analyses of causation. *Noûs*, *36*(2), 276–294.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 225–276). MIT Press.

Hart, H. L. A., & Honoré, T. (1985). *Causation in the law*. OUP Oxford.

Henne, P. (2023). Experimental metaphysics: Causation. In A. M. Bauer & S. Kornmesser (Eds.), *The compact compendium of experimental philosophy*. De Gruyter.

Henne, P., Kulesza, A., Perez, K., & Houcek, A. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition*, *212*.

Henne, P., & O'Neill, K. (2022). Double prevention, causal judgments, and counterfactuals. *Cognitive science*, *46*(5).

Hilton, D. J., McClure, J., & Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, *40*(3), 383–400.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*(11), 587–612.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological review*, *93*(2), 136.

Kominsky, J. F., & Phillips, J. (2019). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive science*, *43*(11), e12792.

Konuk, C., Goodale, M. E., Quillien, T., & Mascarenhas, S. (2023). Plural causes in causal judgment. *Proceedings of the annual meeting of the cognitive science society*, *45*(45).

Krasich, K., O'Neill, K., & De Brigard, F. (2024). Looking at mental images: Eye-tracking mental simulation during retrospective causal judgment. *Cognitive Science*, *48*(3), e13426.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.

Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological review*, *122*(4), 700.

McClure, J., Hilton, D. J., & Sutton, R. M. (2007). Judgments of voluntary and physical causes in causal chains: Probabilistic and social functionalist criteria for attributions. *European journal of social psychology*, *37*(5), 879–901.

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PloS one*, *14*(8), e0219704.

O'Neill, K., Quillien, T., & Henne, P. (2022). A counterfactual model of causal judgment in double prevention. *Conference in computational cognitive neuroscience*.

Pearl, J. (2009). *Causality*. Cambridge university press.

Quillien, T. (2020). When do we think that *X* caused *Y*? *Cognition*, *205*. https://doi.org/10.1016/j.cognition.2020.104410

Quillien, T., & Barlev, M. (2022). Causal judgment in the wild: Evidence from the 2020 u.s. presidential election. *Cognitive Science*, *56*(2). https://doi.org/10.1111/cogs.13101

Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. *Psychological Review*. https://doi.org/10.1037/rev0000428

Quillien, T., Szollosi, A., Bramley, N. R., & Lucas, C. (2023). Causal inference shapes counterfactual plausibility. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Segura, S., Fernandez-Berrocal, P., & Byrne, R. M. (2002). Temporal and causal order effects in thinking about what might have been. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1295–1305.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, *126*(4), 323.

Stan Development Team. (2024). *Stan modeling language users guide and reference manual*. Version 2.35. https://mc-stan.org/

Thanawala, H., & Erb, C. D. (2024). Revisiting causal pluralism: Intention, process, and dependency in cases of double prevention. *Cognition*, *248*, 105786.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, *27*, 1413–1432.

Walsh, C. R., & Byrne, R. M. (2004). Counterfactual thinking: The temporal order effect. *Memory & cognition*, *32*(3), 369–378.

Ziano, I., & Pandelaere, M. (2022). Late-action effect: Heightened counterfactual potency and perceived outcome reversibility make actions closer to a definitive outcome seem more causally impactful [Publisher: Elsevier]. *Journal of Experimental Social Psychology*, *100*, 104290.