Plural causes

Abstract

Causal selection is the process underlying our intuition that an outcome happened *because of* a given event, or that an event is *the cause* of an outcome. When a forest catches fire after a lightning strike, for example, people tend to say that the lightning bolt was the cause of the fire, not mentioning the presence of oxygen in the air, although they are well aware that the latter was no less indispensable for the fire to occur. The extant literature on causal selection has so far operated on the implicit premise that the only relevant variables for causal selection are *individual variables*, corresponding to distinct nodes in the relevant network of causes. Ours is the first systematic study of plural causes in the context of causal selection. First, we establish by means of two behavioral experiments the psychological reality and non-triviality of plural causes, ruling out potential deflationary explanations. Second, we show that state-of-the-art models of causal selection based on counterfactual dependence can be extended to make non-trivial predictions about plural causes consistent with our experimental findings. Third, we show that surprising logical properties of plurals *in natural language interpretation* might be found in causal reasoning with plural causes.

Significance Statement

Why is my child not doing well in school? *Why* are stars moving unexpectedly fast in the outer perimeters of galaxies? Reasoning about causation is central to human intellectual life, because causes provide the answers to *why* questions, both in scientific and everyday discourse. Yet, causal structures are usually too complex to grasp and talk about, so a key task for human understanding is to find ways of *selecting* from a pool of causally active facts the *main cause*. In this work we demonstrate, first, that this task of causal selection doesn't need to pick a single individual event as the main cause, but instead can pick a *plurality* of events. Second, we argue that these plural causes might be represented in human minds with the same resources as pluralities in natural language, suggesting that we can take elements of our theories of natural language as theories of mental representations.

Keywords: causal selection, counterfactual theories of causation, plurals

Introduction

Causal selection is the process underlying our intuition that an outcome happened *because of* a given event, or that an event is *the cause* of an outcome (e.g. Hesslow, 1988; Quillien & Lucas, 2023). Causal selection judgments go further than judgments of *actual causation* (Halpern, 2016; Halpern & Pearl, 2005; Hitchcock, 2001), whereby people merely identify which events can be counted as causes of an outcome. They induce a ranking over these events, singling out some as being more important than others in bringing about the outcome under consideration. When a forest catches fire after a lightning strike, for example, people tend to say that the lightning bolt was the cause of the fire, not mentioning the presence of oxygen in the air, although they are well aware that the latter was no less indispensable for the fire to occur. Causal selection in this sense is crucially distinct from *causal inference*, the problem of learning the relevant causal facts about the world. Causal selection concerns how we judge the relative importance of the many causes of an event, given that we already have a causal model of the situation.

A considerable literature has developed around what factors underly our preference for certain causal explanations of an outcome over others (Icard et al., 2017; Knobe & Fraser, 2008; Lombrozo, 2010; Morris et al., 2019; Quillien & Barlev, 2022). Although theories diverge as to what the drivers of causal selection judgments are, they all agree that the outcome of causal selection judgments depends crucially on the initial pool of candidates under consideration.

Before the lightning bolt can be viewed as *the cause* of the fire, the events *lightning*, *oxygen*, *dry season*, and others must first be flagged by the mind as relevant candidates for causal selection, whose relative importance in bringing about the outcome will be assessed. We argue here that the extant literature on causal selection has had a blind spot regarding that initial pool of candidates: it operates on the implicit premise that the only relevant variables for causal selection are *individual variables*, corresponding to distinct nodes in the relevant network of causes.

Instead, we argue that causal selection judgments can recognize *plural* causes, featuring more than one variable, as when we say that "the dryness of the season and the strength of the wind" caused the spread of the fire. We argue that such plural causes are treated by the mind as candidate explanations on the same footing as the singular causes that compose them. In other words, people engage with such a conjunction of variables as if it were a coherent entity whose impact on the outcome should be evaluated in a wholesale fashion. And the same factors that drive the attractiveness (or lack thereof) of singular-cause explanations drive that of such multivariate causes. This contrasts with an alternative view in which causal selection is at core merely about comparing the importance of "atomic" factors like *dry season* and *strong wind* that can in principle vary independently. On this view, the act of mentioning both in one's explanation would merely be an acknowledgment that one does not have a preference between these two factors, or that each one is important enough that they deserve to be mentioned.

The idea that causal cognition admits causes featuring several variables is not in itself new. In causal inference, researchers have studied how people infer conjunctive causes, that is factors that act in concert to produce an effect (Novick & Cheng, 2004). The notion of a multivariate cause also plays a role in some theories of actual causation (e.g. Halpern, 2015),

and, in a different way, in philosophers' and economists' concept of *collective responsibility* (e.g. Arendt, 1987; Miller, 2001). More specifically related to people's choices of explanations, Lombrozo (2007) and Pacer and Lombrozo (2017) looked at participants' preferences when given a choice between explanations that involved either one or several variables (see also Lucas et al., 2014, for related work with children). They observe that, everything else being equal, human adults prefer explanations that mention fewer unexplained variables (i.e., root causes with no causal parents in the graph). Such studies, however, concern judgments about the ground-truth causal system: which of several causal theories is most likely to be true. By contrast, causal selection, as it is understood here, is primarily concerned with situations in which the ground-truth causal theory — and with it, the variables that count as causes of an outcome — is presumed to be known already. In that context, it asks: which of these causes are most relevant to explain the outcome? To our knowledge, the literature on causal selection judgments has yet to engage with plural explanations.

We present the first systematic study of plural causes in the context of causal selection.[1] This study has three objectives. The first objective is to empirically establish the psychological reality and non-triviality of multivariate causes. In other words, that people assess the value of a plural explanation $A \wedge B$ in a wholesale fashion, treating it as a candidate for causal selection on the same footing as the explanations A and B that mention events contained in it. To that effect we show two things. First, we show that people's judgments about plural causes are sensitive to the prior probabilities of events, a key signature of causal selection judgments. Secondly, and more importantly, we rule out a possible deflationary explanation for plural causes' sensitivity to probabilities: that subjects might formulate a judgment about a plural cause like $A \wedge B$ simply by combining in some direct way their judgments about the importance of the individual events A and B that compose it. In so doing we provide evidence that people treat plurals as full-fledged candidates for causal selection.

The second objective is to show how considering plural causes can expand our understanding of the role of counterfactual reasoning in causal judgments. We show that models of causal selection based on the notion of counterfactual dependence can straightforwardly be extended to make non-trivial predictions about plural causes consistent with our findings. Counterfactual models consider that the causal impact of an event A on an outcome E is a function of the extent to which E depends on A across counterfactual worlds sampled in a certain way. We show that, similarly, people's intuitions as to the causal impact of a plural event $A \wedge B$ is largely captured by the extent to which E depends on $A \wedge B$ across counterfactuals. This cross-validates the counterfactual approach to causal selection by moving to a different class of judgments than the one it was originally developed for.

Our third objective is to explore how some factors other than mere counterfactual dependence as it is currently understood might play a role in subjects' assessment of the contribution of plural causes. In the second experiment of this paper, we find patterns in people's judgments that to our knowledge cannot be explained by any of the existing theories of causal selection judgments. These findings call our attention to the kinds of representations that our causal judgments traffic in. Inspired by the formal semantics of natural language plurals, we speculate that the counterfactual simulation processes people use to evaluate

[1] Our Experiment 1 was presented at the Forty-fifth Annual Meeting of the Cognitive Science Society and published in the society's non-archival proceedings (*1 citation removed for masked review*).

the impact of events might operate at the level of chunks of more than one variable. A computational model formalizing this proposal successfully accounts for the new unexpected results.

Causation and causal selection

Humans are adept at representing the world through a web of causal relations between events. Representing causal relations allows people to make sense of what they observe, make predictions about what's to come, and influence the future in some cases (Chater & Oaksford, 2013; Gerstenberg & Tenenbaum, 2017; Pearl & Mackenzie, 2018; Sloman & Lagnado, 2015).

In the psychological literature, people's causal knowledge is usually modeled through formalisms such as Causal Bayes Nets or Structural Causal Models. These systems represent aspects of the world with variables, causal relations between these variables, and probability distributions (Pearl, 2000). They appear as integral parts of accounts of psychological faculties and functions related to causation, such as causal inference and counterfactual reasoning.

One such causation-related function is causal selection: faced with a complex causal structure, humans will gladly *select* one cause (or, as we will show, more than one) as being more important than others. Moreover, they will assign different scores to different causal variables depending on how they perceive each of those variables as being *the* driver of the observed outcome.

Knowledge of the causal rule in the relevant system is of course one of the main factors determining the explanation humans will favor in causal selection judgments. The other main driver of causal selection is the *normality* attached to events, a notion that combines the extent to which an event abides by moral or conventional rules, and the extent to which it was expected to happen, before it did happen (Icard et al., 2017; Morris et al., 2019; Quillien & Lucas, 2023).

The relationship between the causal rule that entangles events with the outcome, their normality, and causal selection judgment can be complex. In a situation where several different variables are each individually *necessary* for an outcome, people tend to think of the *least expected* variables (the lightning bolt) as *the cause*, and comparatively disregard the importance of the most expected variables (the presence of oxygen), a pattern of judgment known as *abnormal inflation*. The converse tendency is observed in situations where all of the variables considered are each individually *sufficient* for the outcome to occur. In this case, people tend instead to think of the most normal events as the most important causes of the outcome (Icard et al., 2017).

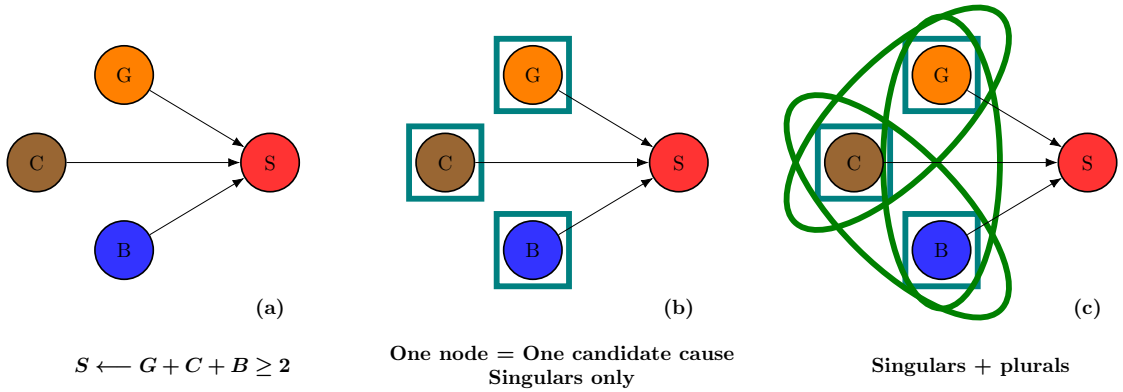
Defining the candidates for causal selection

Causal selection is determined by an amalgam of the system's underlying causal rule and the normality of events. Thus, a standard procedure for formulating theories about participants' causal selection judgments starts by building a causal model that formalizes their causal knowledge of the system.

Suppose for example that I get a stomachache shortly after having eaten a piece of Gouda cheese and a plate of pudding containing chocolate cake and blueberry pie. A causal model of this situation would feature one variable for each of the causes of my stomachache (i.e. one variable each for "eating the Gouda cheese," "eating the chocolate cake," and "eating

Figure 1

A causal model for the relations between various dishes and my stomachache. **(a)** *I develop a stomachache if and only if I eat two kinds of pudding or more.* **(b)** *The standard implicit assumption in the literature is that only single variables are candidates for causal selection.* **(c)** *We propose instead that causal judgments can also target plurals, for example pairs of variables.*



the blueberry pie") as well as a variable for the effect ("having a stomachache"). The model also specifies a functional relationship between the variables, for example representing the fact that one develops stomach issues after eating too much, as schematized in Figure 1a.

As illustrated by this representational format, it is natural to think of the candidates for causal selection as particular realizations of the individual variables. If an individual equipped with the causal knowledge encapsulated by the model in Figure 1a wonders what *the cause* of their stomachache was, it may seem like they have to make a choice between the variables G, C, and B. This would directly identify the candidates for causal selection as the individual moving parts of the causal model, as represented in Figure 1b.

A striking feature of the psychological literature on causal selection is indeed that causal selection judgments are only ever queried at the level of singular variables (Kominsky et al., 2015; Morris et al., 2019; Quillien & Barlev, 2022; Quillien & Lucas, 2023; Sytsma, 2020). Kinney and Lombrozo (2024) deserve an honorable mention in this connection however, since they compared participants' preferences for causal generics ("X causes Y") mentioning one vs. several variables. But their work was on type causation, while here we discuss token (actual) causation.

Concretely, when experimental participants are presented with a situation where an outcome depends on three different events A, B, or C, they are never asked to what extent a *plural* event like $A \wedge B$ can be considered the cause of the outcome. Intuitively though, causal explanations that mention combinations of variables can also be appealing. In our example above, saying that I got a stomachache "because I ate the entire dessert plate" might appear to be a better explanation than either "because I ate the chocolate cake" or "because I ate the blueberry pie" each on its own.

Note that allowing for many variables to feature in causal explanations does not elim-

inate the need for causal selection: one might want to mention several causes of an event without mentioning *all* of them. For example, one might think that "because I ate the blueberry pie" is a better explanation for my stomachache than "because I ate the entire dessert plate" if for example I eat chocolate cake at every meal, but add a blueberry pie on top of it only exceptionally. Ultimately, the best candidates for causal selection are those causes that participants see as most *crucial* in bringing about the outcome, whether these be singular or plural, and in principle we can only know what the best causal explanations are after considering the entire set of possible candidates, including plural causes, as illustrated in Figure 1c.

Counterfactual theories

To properly argue the point above, we first need to spell out what it means for a cause to be of a more or less crucial importance in bringing about an outcome. The notion we will rely on throughout this paper is rooted in counterfactual theories of causal selection (Icard et al., 2017; Quillien & Lucas, 2023).

Counterfactual theories of causal cognition in general build on the premise that humans represent causal relations between variables in terms of counterfactual dependence (Gerstenberg & Tenenbaum, 2017; Halpern & Pearl, 2005; Krasich et al., 2024; Lewis, 1973; Woodward, 2003, 2006). The notion that "C caused E" is taken to be roughly equivalent to the notion that "had C not happened, E would not have happened either."

In the case of causal selection judgments, this is enriched by evaluating counterfactual dependence not just once but across many possible worlds, asking how robustly E depends on C when background conditions vary. Of particular relevance to this evaluation will be the possible worlds that are most *normal*, or *closest to the actual world* in which we are to select a cause (Lewis, 1973). Evaluating counterfactual dependence in these worlds is what allows a causal selection judgment to provide explanations that are not just relevant to the situation under consideration, but also generalizable to other contexts (Hitchcock, 2012; Lombrozo, 2010).

We will limit our discussion in this article to two counterfactual theories (and accompanying models) of causal selection judgment that (1) have been stated in full mathematical rigor and (2) have been submitted to experimental scrutiny, the Necessity and Sufficiency Model (Icard et al., 2017, NSM) and the Counterfactual Effect-Size Model (Quillien & Lucas, 2023, CESM). We chose to focus on these two theories because of their good track record in predicting participants' causal selection judgments across a wide variety of tasks (Gerstenberg & Icard, 2020; Gill et al., 2022; Henne et al., 2019, 2021; Kirfel et al., 2021; Kominsky & Phillips, 2019; Morris et al., 2019; O'Neill et al., 2022; O'Neill et al., 2025; Quillien & Barlev, 2022).

The two theories see causal selection as a two-step process. The first step is identical across theories, the second divergent. The procedure is as follows.

First, randomly sample a large number of counterfactual worlds. The sampling process operates at the level of the individual exogenous variables of the relevant causal model, that is the variables that have no parent in the causal graph. Across worlds, each of these variables is sampled with a frequency that is a function of two elements:

1. Its value in the actual world. Given a causal model with a set of exogenous variables E, and a valuation function $[\![\cdot]\!]^w$ that maps each variable in E to the value it has in a world

of evaluation w, we can define a special world constant $w_@$ designating the actual world, that is the set of circumstances that in fact took place. The model includes a stability parameter s taking a value between 0 and 1, such that in every counterfactual world $w_i \in \{w_1, \ldots, w_n\}$ that it samples, each variable in E will have in w_i the same value that it has in the actual world $w_@$, with probability s (see Lucas & Kemp, 2015; Quillien et al., 2023). This parameter is not present in the original version of the Necessity and Sufficiency Model, having been introduced in models of causal selection by Quillien and Lucas (2023). It can however be straightforwardly added to it, as we will do in this article.

2. The variable's prior probability of occurrence. When a variable's value is not directly mapped to its actual world value, as will happen with probability $1-s$, it is resampled from its prior probability distribution. This is where an event's sampling propensity (and from there, its causal score) gets to be sensitive to the normality attached to that event.

Consider the example of the causal system presented above relating variables G, C, and B to my stomachache. In the actual world, the variable G that encodes my eating Gouda cheese has value 1, meaning that event actually took place. As a result, when I sample counterfactuals to the actual world (as in Figure 2), I will with a probability s automatically represent myself eating cheese also in each of these worlds, as depicted in the left side of the tree in Figure 2a. In the worlds where I don't do that (the right side of the decision tree), the variable will be resampled from the prior probability on that event. Suppose for example that my eating cheese is a rather exceptional occurrence, such that $P(G) = 0.1$. In this case, I am much more likely to travel along the rightmost sub-branch of the tree in Figure 2a, and simulate worlds in which I didn't eat Gouda cheese than if I were accustomed to the fact and ascribed a higher prior probability to the event.

All in all, the sampling propensity (Icard, 2016) of any given exogenous variable V can be reconstructed as a function of the stability parameter s and that variable's prior probability $P(V)$, following the equation below, in the special case of interest here (binary variables, registering whether an event happens or not).

$$SP(V) = s \cdot [\![V]\!]^{w_@} + (1-s) \cdot P(V)$$

Once the exogenous variables of the system have been sampled in this way, one can simulate the outcome, which follows from the variables via the causal rule underlying this particular system, as in Figure 2b.

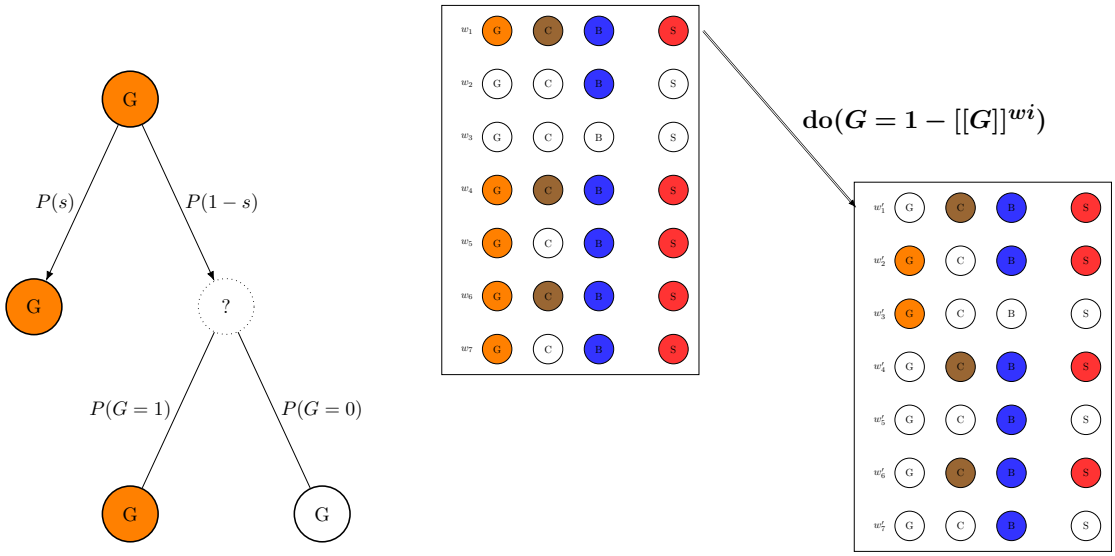
Second, compute the causal impact of a given variable V across those counterfactual worlds. The precise way to measure this impact is different in the two theories under consideration. In the NSM, causal impact is scored as the weighted sum of the following two factors.

1. A Necessity score: In each world w in which $[\![V]\!]^w \neq [\![V]\!]^{w_@}$, sample the outcome O from a probability distribution $P^\nu(O)$, where the value of each variable $V_j \in E$ other than V is switched to its actual world value $[\![V_j]\!]^{w_@}$. Then, count one point for the necessity score if the value of the outcome in the resulting world is different from the one that it had in the actual world (i.e. $P^\nu(O \neq [\![O]\!]^{w_@})$), and zero points otherwise.

2. A Sufficiency score: For every world w in which $[\![V]\!]^w = [\![V]\!]^{w_@}$, sample the outcome from the probability of Sufficiency $P^\sigma_{V=[\![V]\!]^{w_@}}(O)$ of the event $V = [\![V]\!]^{w_@}$ for the outcome O. There is more than one way to define $P^\sigma_{V=[\![V]\!]^{w_@}}$, but they make extremely similar predictions.

Figure 2

Sampling counterfactual worlds.



(a) *Decision tree representing how the parameter s anchors the sampling of a variable to its real-world value, making variables more likely to be the way they were in the real world (following Lucas & Kemp, 2015).*

(b) *Right: representation of a sample of seven worlds generated with the sampling process illustrated in (a). In the real world (top line w_1), all events are present. But some (like C) had a lower prior probability than others (like B), hence they feature less often across counterfactuals $w_2 \ldots w_7$; left: the same worlds, after an intervention flipping the value of G, with the resulting consequence on the outcome S.*

The definition that turned out to have the best fit with the data from our experiments is the following.

$$P^{\sigma}_{V=[\![V]\!]^{w_@}}(O) = SP(O \mid do(V = [\![V]\!]^{w_@}), \neg[\![V]\!]^{w_@}, \neg O)$$

From here, count one point for the sufficiency score if the value of the outcome thus sampled in w is the same as the value of the outcome in the actual world, and zero points otherwise.

Then, divide the number of points scored this way by the total count of worlds sampled. The dynamics of necessity and sufficiency scoring make it such that the necessity score is all the more important when the prior probability of an event is low, making it more likely to switch value across counterfactuals, whereas the sufficiency score is all the more important when this prior probability is high.

In the CESM, causal impact is computed using the same process in every world w_i, as follows.

1. Switch the value of the variable V to a new, randomly sampled value. Then reevaluate the outcome in the new world w_i' where the value was switched. A representation of this resampling process is given in Figure 2b. The impact $K(V \rightarrow O_i, w_i)$ of V in the world w_i is

then evaluated as

$$K(V \to O_i, w_i) = \frac{\Delta O}{\Delta V} = \frac{[\![O]\!]^{w_i} - [\![O]\!]^{w_i'}}{[\![V]\!]^{w_i} - [\![V]\!]^{w_i'}}.$$

This equation can be glossed as follows. Whenever the outcome is switched in the same direction as the target variable (both from 1 to 0, or both from 0 to 1), V scores a point; when the outcome is unaffected, it scores none. When it moves in the opposite direction, it scores a negative point.

2. The causal impact is then normalized by the ratio of the standard deviations $\frac{\sigma_O}{\sigma_V}$, and averaged across worlds to get the causal impact score $K(V \to O, w_@)$ of the target variable for the target outcome in the actual world.

In simple causal structures like the ones we deal with in this article, the causal impact of V is equivalent to the correlation coefficient between V and O across counterfactuals sampled at the first step.

Extending causal selection to plurals

We hypothesize that humans consider plural causes as *bona fide* candidates for causal selection and that they assess their impact by tracking how an outcome counterfactually depends on them just like they would do for single-variable causes.

This hypothesis produces testable predictions. For one thing, we would expect plural explanations to be sensitive to prior probabilities in the same way that singular ones are, since prior probabilities shape the counterfactual dependence between causes and outcomes. But this is not enough to show that people assess the impact of plurals/conjunctions as such. An alternative view could indeed contend that people only ever have direct intuitions about the causal responsibility of the *individual* variables in their causal model, but can still make judgments about a plural cause say by adding up or averaging the individual causal strengths of its constituent variables.

For example, to compute how much they agree that "eating the chocolate cake and the blueberry pie caused the stomachache," people might first compute their agreement with "eating the chocolate cake caused the stomachache," "eating the blueberry pie caused the stomachache," and so on. Then they might somehow aggregate the causal strength of each individual variable. We will call this the *linear combination* hypothesis. Under this hypothesis, each variable makes a fixed contribution to any plural it is part of, regardless of what other variables it is combined with: the contribution of "blueberry pie" to "blueberry pie & chocolate cake" would be the same as its contribution to "blueberry pie & cheesecake." There are no interaction effects — no causal relevance that emerges only from the combination. This hypothesis is deflationary with respect to the psychological reality of plural causes in that it holds that people can make plural-cause judgments when prompted to do so, but they cobble them together from more primitive representations of causal strength at the level of individual variables. In such a view, the internal computations relevant to causal selection ultimately only attend to the individual components of the system. The sensitivity of plural causes to prior probabilities is merely a surface effect: one sees a plural $A \wedge B$ as a better explanation than $B \wedge C$ merely because it comprises a variable A with a better counterfactual dependence profile than the variable C.

In contrast, we consider the possibility that plural causal judgments are the output of a *holistic* computation. Under this possibility, the same cognitive process that allows people to formulate causal judgments about singular variables is deployed at the level of combinations of variables, yielding quantities that can on occasion diverge significantly and non-linearly from the causal judgments for constituent variables. We consider in detail how such a hypothesis could be implemented in models of causal selection in the next section. But the general idea can be explained rather simply: to assess the causal importance of a plural event $A \wedge B$ is to look at the causal impact that this *compound event* has on the outcome of interest, using the same measures of causal impact that were detailed above for singular variables.

This holistic computation will sometimes lead to different predictions than the hypothesis that consists in simply computing the impact of A, of B, and then combining them. In order to tease apart these two hypotheses in the case of causal selection then, we need to identify contexts where they make different predictions about causal selection judgments. This is what we do in our first experiment. It establishes that plural causes are real psychological entities: to say that "A and B" caused an outcome is not merely to say that variables A and B are each individually relevant. It is to say that the outcome counterfactually depends on variables A, B taken together as a cluster, more so than on alternative candidates. Having established this, in Experiment 2 we move on to explore how this clustering of variables interacts with the groupings implicit in the causal function that relate an outcome to all of its causes, such as when it accepts two possible sufficient conditions $(A \wedge B)$ and $(C \wedge D)$.

Experiment 1

Our first experiment has the following goals.[2] First, if plural causes are processed as genuine causes by the mind, factors that are known to affect causal selection judgments should influence judgments about a plural cause. In particular, the probability of an event is known to affect judgments about whether that event caused an outcome (Morris et al., 2019). If plural causes are genuine candidates for causal selection, then we would expect analogous patterns to apply to them: varying the probability of events should affect causal judgments about whether a conjunction of these events is seen as causing the outcome. Second, we aim to rule out the deflationary *linear combination* account of the impact of probability on participants' judgments. Evidence of non-linearity in causal judgments would constitute stronger evidence for the psychological reality of plural causes in human causal selection. We design a situation where both the CESM and the NSM predict that the causal strength of plural variables will not be a linear combination of the score of individual variables. We compare their predictions to those of a null-hypothesis model that tries to predict the score of plural causes as a linear combination of the scores of individual variables.

[2] This study was reported at the Forty-fifth Annual Meeting of the Cognitive Science Society and published in the society's non-archival conference proceedings (*1 citation removed for masked review*). Our writing in this section borrows directly from this preliminary report.

Methods

Design and materials

We adapted a paradigm developed by Quillien and Lucas (2023). Participants made judgments about a game of chance, in which one randomly draws balls from a set of urns, and wins by getting enough colored balls (see Figure 3 for illustrations). Participants observed a fictitious player draw a colored ball from each of three urns (labeled A, B, and C) and win the game as a result. Then they were asked to make a causal judgment about each singular cause (e.g. whether getting a colored ball from urn A caused the player to win the game), and about each pair of causes (e.g. whether getting a colored ball from urns A and B caused the player to win the game). For exploratory purposes, we also asked participants to make a causal judgment about the triplet (getting a colored ball from A, B, and C). We manipulated the prior probability of each outcome within participants by varying the proportion of colored balls in each urn, with probabilities of 0.05, 0.5, and 0.95 (Figure 3). We will refer to the three different urns as the low, intermediate, and high urns, respectively. The rule of the game, which was directly revealed to the participant at the outset, was that the player wins if they get two colored balls or more. This corresponds to the causal model below.

$$\text{WIN} := A + B + C \geq 2$$

Predictions

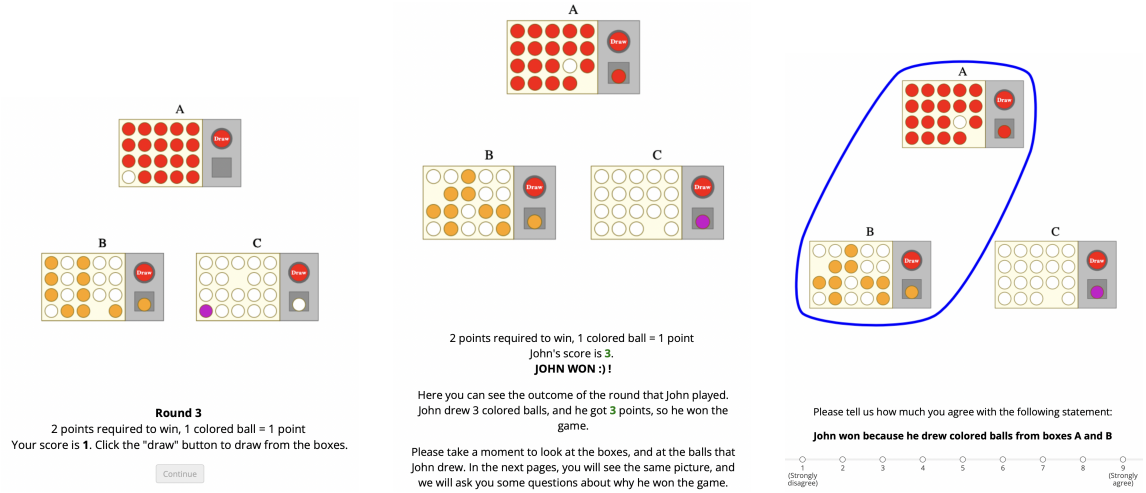
This paradigm provides a context where the linear and the holistic extensions of the models we outlined above make clearly different predictions. The CESM predicts that participants' singular causal-strength estimates should follow a particular ranking: intermediate $>$ low $>$ high, for any value of the s parameter. This is because, across possible counterfactual alternatives to what happened, there is a high correlation between getting a colored ball from the intermediate probability urn and winning the game. These predictions partially match participants' judgments collected in the previous iteration of this paradigm run by Quillien and Lucas (2023), where judgments were collected for singular variables only, and in which participants' responses followed the ranking: intermediate probability urn $>$ low $\approx$ high. There was no significant difference between the ratings for the low probability and the high probability urn.

If we consider an extension of the CESM, linearly combining the predictions for individual causes, or alternatively if we simply combine participants' judgments on individual variables, we would predict that participants should consider that the pair low & intermediate should have a causal strength greater than or equal to that of the pair high & intermediate, because the singular low has higher causal strength than high.

In contrast, if participants judge the causal strength of plurals via a holistic computation, they should rate the pair intermediate & high as highest. For across possible counterfactuals there is a high correlation between getting a colored ball from these two urns and winning the game. Intuitively, since drawing a colored ball from the low-probability urn is rare, and given that at least two balls are needed to win, most worlds where the player wins the game will be worlds in which they do so by getting a colored ball from the intermediate and high urns. This prediction is true for any value of the s parameter in the holistic version of

Figure 3

The three phases of Experiment 1



(a) *First, participants played 10 rounds of the game to get familiar with the rule and the underlying probabilities. The rounds were pseudo-randomized so as to make sure they reflected the probabilities associated with each urn.*

(b) *They then observed the outcome of a round of the game played by a different player named "John." In this experiment, John would draw a colored ball from each of the three urns and therefore win in an overdetermined fashion (getting three points where he needed only two).*

(c) *We queried their judgments about every singular and plural variable. Causes mentioned in the sentence in boldface were also visually highlighted by circling the corresponding urns on the screen.*

the CESM. It is also true for the holistic version of the NSM, although in that case it reverses the ranking that the NSM expects for singulars.

Procedure

Participants first completed ten rounds of the game, presented with urns as in Figure 3a. We pseudo-randomized the draws in such a way as to get participants to internalize the probabilities associated with each urn and how they connected to the outcome. Then participants saw the outcome of a round of the game played by another (fictitious) player, who drew a colored ball from all three urns, thereby winning with 3 points, as in Figure 3b. They were asked to rate the causal strength of each individual draw, as well as that of every combination of two or three draws for the winning outcome, on a Likert scale from 1 to 9 (strongly disagree to strongly agree), following standard practice in the causal selection literature (Icard et al., 2017; Morris et al., 2019), as in Figure 3c. For the singulars, participants were asked to rate their agreement with the statement "John won because he drew a colored ball from box [urn]." For the plurals, they rated their agreement with "John won because he drew colored balls from boxes [urn1] and [urn2]." Each question was shown on a separate page, next to the urns that displayed the outcome of the fictitious player's draw. The letters indexing the urns,

as well as the colors of the balls, were randomized across participants but were kept the same for all trials within participants. Half of the participants were asked about the singulars first, and then about the pairs. The other half were asked about the pairs first, and then about the singulars. All participants were asked about the triplet at the very end. Within one class of questions (singulars vs. plurals) we randomized the order of presentation of questions. Finally, participants completed a brief demographic questionnaire and were redirected to Prolific for payment. We coded the experiment in the `jsPsych` library (De Leeuw, 2015), with custom plugins for displaying urns developed in our lab.

Participants

We recruited 400 participants from all English-speaking countries from Prolific. This sample size was inspired by the one used by Quillien and Lucas (2023), who used a comparable sample size (290 participants) for a study with similar design. We excluded from subsequent analysis 44 participants who failed to answer either of two elementary comprehension questions that checked their understanding of the rules of the game, leaving a total of 356 participants for analysis.

Transparency and openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. All data, analysis code, and research materials are available at https://osf.io/43m5d/?view_only=9843a8609a1b49e9bd630231984c588c. Data were analyzed using `R` (R Core Team, 2013) version 4.3.3 and the `tidyverse` package collection (Wickham et al., 2019), version 2.0.0. All studies we report in this article received ethics approval by the *Comité d'évaluation de l'éthique de l'INSERM*, under research protocol *Le langage et les capacités cognitives connexes*. All studies were conducted entirely in English. We did not preregister the studies.

Results

We first report analyses using standard statistical tests. Then we report the fit of computational models of causal judgment.

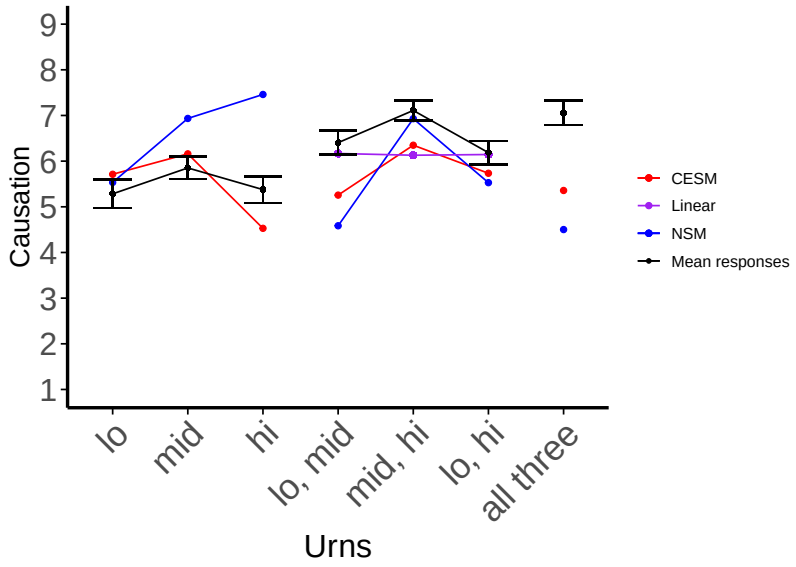
Basic results patterns

Prior probability affects both singular and plural causal judgments. Results are plotted in Figure 4. We ran two two-factor repeated-measure ANOVAs, one for each main type of cause queried (*singulars* and *pairs*), using urn probabilities and order of presentation as predictor variables, and participants' responses as the dependent variable. Results are in Tables 1 and 2. There was a main effect of prior probability on participants' causal judgments, for singular as well as for plural causes ($p < 0.02$ in both cases, see the tables for test statistics), consistent with our expectation that participants' judgments of plural causes should be sensitive to probabilities just like for other actual-cause judgments.

The order of presentation/querying singular and plural selection judgments also had a significant effect ($p < 0.001$) on the ratings for singulars: singular causal judgments were lower when presented after the plurals. There was however no interaction effect between urn probability and order of presentation, suggesting that the impact of probability on causal

Figure 4

Mean ratings by question type, along with predictions for each theory under consideration. The error bars represent the standard error of the mean. The linear combination *theory (purple) predicts that the score of the* low & intermediate *and* intermediate & high *pairs should be equivalent, when in fact we see a significant difference between them, in accordance with the* holistic *versions of the two counterfactual models: the Counterfactual Effect Size Model (Quillien & Lucas, 2023) (in red on the plot) and the Necessity and Sufficiency Model (Icard et al., 2017) (in blue).*



estimates did not vary depending on the order in which questions were asked. Therefore we drop this variable (order of presentation) from later analyses.

The causal strength of plural causes is not a linear combination of the causal strength of individual variables. The pattern of responses for singular variables replicated the patterns obtained by Quillien and Lucas (2023). Judgments for the intermediate urn were higher than judgments for the low urn, $t(315.41) = -2.70$, $p = 0.007$, and the high urn, $t(325.85) = -2.08$, $p = 0.038$. The difference between the low and high urns was not significant, $t(350.59) = -0.63$, $p = 0.53$.

We can use these results to test the *linear combination* hypothesis, according to which participants derive their plural-cause strength estimates by adding up or averaging their estimates for the individual variables that compose a given plural. If this were correct, participants should give the same causal strength estimate for the two plural causes low & intermediate and intermediate & high, since their estimates for the singular causes low and high are not significantly different from each other. By contrast, both the holistic CESM and the holistic NSM predict a sharp difference between these two kinds of plurals, with the plural cause intermediate & high being rated higher (Figure 4).

Consistent with the holistic CESM, judgments about the intermediate & high pair were higher than those for the low & intermediate pair, $t(355) = -4.67$, $p < 0.001$, and higher

Table 1

ANOVA for singular causal selection judgments, predicting urn ratings from urn probabilities and urn order of presentation.

Factors	Mean Sq	F-score	p-value
Probabilities of the urns	32.98	4.492	< 0.012
Order of presentation	138.81	18.904	< 0.0001
Probabilities:order	3.18	0.433	> 0.64

Table 2

ANOVA for pair causal selection judgments.

Factors	Mean Sq	F-score	p-value
Probabilities of the urns	83.02	14.646	< 0.000001
Order of presentation	21.75	3.837	> 0.05
Probabilities:order	2.30	0.406	> 0.66

than for the low & high pair, $t(355) = 6.858$, $p < 0.001$. In a slight deviation from the CESM and NSM's predictions however, judgments for the low & intermediate pair were higher than for the low & high pair, $t(355) = 2.3691$, $p = 0.02$ (Figure 4).

We conducted two more analyses to rule out the linear combination model. First, we ran a one-way repeated-measure ANOVA, predicting judgments for the pairs (low & intermediate, intermediate & high, and low & high) from judgments for the singulars (low, intermediate, and high), as well as their interactions, as within-participant factors. Each plural pair was regressed only on the values of the two singulars that comprised it.

The linear combination theory predicts that there should be no significant interaction: a participant's causal judgment for a given *singular* variable should have the same impact on every plural cause in which it features. One's estimate for the singular intermediate, for example, should have an equal impact on one's estimate for intermediate & high and for low & intermediate.

We find evidence against the linear combination theory (Table 3). There was a significant interaction between the intermediate and high urns, $p = 0.004$. In addition, the main effects of the singular judgments were not significant, for all but the low urn.

Second, we fitted linear multilevel regression models on participants' responses for pairs. Specifically, we compared the predictive performance of two different models on participants' plural-cause estimates. The first one used as predictor the average of the two singular-cause estimates for the variables contained in a given plural (computed on a per-participant basis), plus a random intercept. The second model also included the question asked (that is, the specific plural being queried) as predictor. A likelihood-ratio test shows that adding question as a predictor significantly improves the fit of the model, $\chi^2(5) = 27.53$, $p < 0.001$ (Table 4). Again, this is inconsistent with the linear combination account.

Table 3

Results of the ANOVA: estimate for pairs ~ est. for singular-1 × est. for singular-2.

Factors	Mean Sq	*F*-value	*p*-value
low	100.69	17.743	< 0.00001
intermediate	20.80	3.664	0.05586
high	15.93	2.808	0.094
low:intermediate	6.64	1.169	0.2798
low:high	0.41	0.073	0.7872
intermediate:high	46.64	8.219	0.00423

Table 4

Comparison between two models: the linear combination model of plurals (means of singulars + intercept), and the means of singulars + question model.

Models	LogLik	Df	χ^2	*p*-value	BIC
Means sing	-891.89	3			1804.709
+ Question	-878.13	5	27.53	< 0.00001	1791.126

Computational modeling

We computed the predictions of two recent counterfactual models of causal selection, the Counterfactual Effect Size Model (Quillien & Lucas, 2023) and the Necessity and Sufficiency Model (Icard et al., 2017), presented in the introduction. Our implementation follows the one given by Quillien and Lucas (2023).

For each question we report on below, we generated causal judgments for the CESM using a process of counterfactual sampling. We generated predictions for the CESM by simulating 10^5 possible rounds of the game according to the rule, what was the case in the situation described to participants, and the sampling model described by the CESM. We computed CESM judgments for an event as the correlation between that event (for instance, whether the player draws a colored ball from urn *A*) and the outcome of the game (whether the player wins the game), across simulations. We computed NSM judgments analytically, as the sum of the variables' sufficiency and necessity scores across worlds.

We fit the value of the stability parameter s for both models by finding the value of s that results in the best fit between model judgments and average participant judgments across all seven questions. We quantified model fit by looking at the likelihood of mean answers per question under a normal distribution centered on the model's predictions, with a standard deviation fitted across questions.

We identified the best fit value via a grid search, exploring a wide range of values for the parameter s, crossed with different values for a scaling parameter γ (applied to a model's predictions as an exponent $prediction^{\gamma}$). The point of γ was to avoid situations where one model would systematically overshoot or undershoot actual participant answers, as the models are not meant to predict the exact value of participants' judgments, but only the relative

Table 5

Model comparison for Experiment 1, excluding the triple. The AIC and BIC values are computed for mixed effects models, including group and a random effect for participants.

Model	AIC	BIC	Cor.
CESM	9929.31	9957.65	0.54
NSM	9962.06	9990.4	0.24
	Considering only the pairs		
CESM	4692.11	4716.978	0.83
NSM	4674.355	4699.222	0.80
Empirical average	4692.942	4717.81	-0.65

difference between one variable and the next. Our technique here was analogous to that of Griffiths and Tenenbaum (2005).

For the CESM, the best fitting value was $s = 0.89$, with $\gamma = 0.26$. For the NSM, the best fitting value was $s = 0.71$, with $\gamma = 2.93$.

In our implementation, to assess the causal strength of plural causes a model assumes that people compute the causal strength of the conjunction of all variables contained within that plural. For instance, the CESM computes the causal strength of low & high by computing the correlation between the compound binary variable low $\wedge$ high (which has value 1 if both low and high have value 1, and 0 otherwise) and the outcome.

The predictions of the models are plotted in Figure 4. Table 5 details the comparison. Overall, the CESM's predictions had the best fit to human judgments in this experiment, although the NSM had the best fit when models were compared on pairs of variables only. We also compared the models' performance on the judgments for pairs to a null model that used as predictor for each pair the average of mean human judgments for each singular variable contained within a given pair, as plotted in Figure 4. Both counterfactual models proved significantly better than this linear predictor (Table 5).

Discussion

We find evidence that, when people make a judgment about whether events A and B caused an outcome, their judgments track the correlation between the conjunction of A and B and the outcome, across counterfactuals. Concretely, in our experiment, winning the game is in general strongly associated with getting a ball from both the intermediate- and high-probability urns, and people judged that combination of events to be highly causal. Importantly, this effect is inconsistent with a simpler account, according to which people's judgments about plurals are cobbled together from their causal intuitions about each individual variable in the plural.

Judgments about plural causes are affected by the prior probability of their constituent variables, but cannot be derived from the causal strength of these individual variables. As such, our results are in general consistent with the predictions of simple extensions of recent counterfactual models of causal selection (Icard et al., 2017; Quillien, 2020; Quillien & Lucas,

2023), augmented with the assumption that people judge plural causes in a holistic manner.

At the same time, these findings raise new questions about the psychology of causation. Presently we highlight two of these questions, which we investigate in Experiment 2.

First, participants in this study found the plurals overall more appealing than the singulars, a tendency which the counterfactual models we considered did not capture. Participants might have felt that plurals provided more exhaustive descriptions of the event: they give more complete information about what happened, in addition to why it happened. We also find that this effect is accentuated when singulars are presented after plurals. Making judgments about plurals first might highlight to participants the descriptive incompleteness of singulars. This finding suggests an interesting tension between two potential desiderata of causal judgment: highlighting the variables that were most causally important to the outcome, and providing an exhaustive list of the causal factors. If so, this calls for an investigation into the relative importance of these two pressures in participants' causal selection judgments. When, for example, adding a variable weakens the counterfactual dependence profile of the resulting plural, such as when the plural doesn't explain the outcome appreciably better than one of the singular variables within it, will participants still show a preference for plurals, on account of their greater completeness?

Second, a notable feature of this first experiment is that the causal structure used a simple additive rule (i.e. the player wins the game if their score is above a certain threshold). As such, there is a sense in which the variables each have an independent incremental causal effect on the outcome.

What will participants' plural-cause judgments look like in a causal structure where some conjunctions of events directly feature as such in the causal model that generates the outcome? Consider for example the causal rule $(A \wedge B) \vee C$. Here the urns A and B are specifically connected in the logical structure. Generalizing somewhat, our question here is: when an outcome specifically depends on the joint occurrence of A and B, should that make the plural cause $A \wedge B$ a more natural causal explanation than a potential alternative $A \wedge C$, even if C also makes an important contribution to the outcome?

Experiment 2

Experiment 1 established the psychological reality and relevance of plural causes for causal selection judgments. It provides evidence that, in order to assess the value of a conjunctive explanation like "E happened because of A and B," people would track the way in which the group constituted by variables A, B contributes to E across counterfactuals. Building on these findings, Experiment 2 expands our exploration of plural causes by looking at a richer causal structure. Here, two urns contain purple balls, and two urns contain orange balls. The player can win the game by getting either two purple balls or two orange balls, where "or" is meant inclusively. Formally, winning can be triggered by either of two distinct sufficient conditions $A \wedge B$ and $C \wedge D$, each a conjunction of two variables. This corresponds to the rule

$$\text{WIN} := (A \wedge B) \vee (C \wedge D). \tag{1}$$

The point of moving to such a richer rule is fourfold.

First, we provide additional evidence against deflationary interpretations of plural causes. We give more examples of situations in which people's plural-cause judgments cannot

be straightforwardly derived from a linear combination of their singular-cause judgments, to confirm the results obtained in the first experiment.

Second, we explore whether there is a robust bias toward preferring causes that contain more variables. In Experiment 1, participants gave overall stronger scores to plural causes than singular ones. Experiment 2 investigates whether this pattern always holds.

Third, we study how the groupings involved in plural explanations interact with those implicit in the causal rule linking variables to their outcome. Intuitively, given knowledge of the rule (1), and of the value of variables A, B, C, D in some (actual or counterfactual) world, in order to assess that a round is won one might want to check each sufficient condition separately: whether $A \wedge B$ is true, and whether $C \wedge D$ is true. The two mental operations might run in parallel; but what matters is that each pair is attended to separately, therefore inducing its own implicit grouping, which in this case follows the boundaries of the logical disjunction.

This raises the question of how such implicit groupings will affect the way in which subjects like to group variables when they engage in plural explanations. An interesting possibility to explore is that people might be evaluating the contribution of causes by looking at how much they contribute to one of these two paths, rather than directly looking at how they correlate with the outcome itself. To foreshadow our results, we did find some evidence in favor of such a view, by showing that people tend to dislike plural explanations that "cross" the disjunction, in a way that goes above and beyond what is predicted by the counterfactual dependence profile of these variables alone.

Fourth, we explore participants' judgments in situations where they have to explain a *negative* outcome. In the context of our experiment, this amounts to explaining a loss. We call losing a *negative* outcome here, not in the sense of an undesirable outcome, but in the sense that subjects have only been positively instructed in the sufficient conditions for winning. The conditions for losing are merely contained *implicitly* in these instructions, as the negation of winning conditions. As we detail presently, this contrast opens interesting possibilities about the representations participants might have for losing conditions.

Finally, Experiment 2 was also designed to collect many more data points per participant, increasing our statistical power compared to Experiment 1. We ask each participant about the outcome of four possible rounds of the game (as opposed to just one outcome in Experiment 1), collecting a total of 36 causal judgments per participant.

Negative outcomes and plurals

Responsibility attributions for negative outcomes are relatively understudied. Part of the reason for this might come from the expectation that the shift from positive to negative should be trivial, especially for binary variables. The same processes by which we assign responsibility to wins can be repurposed for losses, simply by moving the target. Erstwhile "wins" now count as losses and "losses" become wins, as far as assigning credit to causes goes. This would be true if we assume that the processes by which counterfactual outcomes are determined amount to consulting a model that already matches each assignment of values to causes to an outcome. In which case, explaining a loss would simply amount to tracking how different causes co-vary with the *classical logic negation* of winning conditions across counterfactuals.

But available evidence suggests a more complex picture. Recently, in a study of *ex ante* responsibility judgments (that is, judging the importance of various causes before any outcome has effectively occurred or the value of any causal variable is known), Gerstenberg et al. (2023) observe that subjects' estimates are better captured by a measure that tracks a variable's contribution to wins than one that tracks its contribution to losses (or some hybrid of the two). This suggests that positive outcomes are taken to be the explananda by default for responsibility judgments. Knowing what it takes to win, one may go about checking whether these conditions are satisfied in each counterfactual, and credit causes as a function of how they contribute to satisfying these conditions.

But when people try to explain a loss, they might not simply reconstruct the losing conditions by computing the classical logical negation of winning conditions. A case in point here is provided by a series of experiments by Gerstenberg and Icard (2020), who looked at causal selection judgments in a billiard-ball setting involving simple conjunctive or disjunctive rules. The researchers collected subjects' estimates in positive cases where two events A and B happened, and in negative cases where both A and B failed to happen. As they noted, judgments for negative outcomes in the *disjunctive* cases where $E := A \vee B$ were neatly captured by treating negative outcomes as the classical negation $\neg E := \neg A \wedge \neg B$. But the same strategy did not work in the conjunctive case where $E := A \wedge B$. As we will see shortly, we found related results in our Experiment 2, where subjects' judgments did not fit the pattern expected by the classical interpretation of negated conjunctions.

To foreshadow these results, we instead find that many patterns of judgments in that condition can be captured by assuming that participants are checking whether the player loses the game in a given counterfactual simulation by using the formula

$$\text{LOSS} := \neg A \wedge \neg B \wedge \neg C \wedge \neg D.$$

Note that this is not the correct equation one would obtain by applying classical logical negation to the equation for winning the game. Nonetheless, a counterfactual model that assumes that people simulate counterfactual possibilities in this *non-classical* way provides a very good fit to the judgments collected in the losing rounds of our experiment. Of course this finding raises the question: why would subjects rate explanations as if they endorsed a representation of the losing conditions that is incorrect? And also: why (as suggested by the data from Gerstenberg & Icard, 2020) would they only engage in such non-classical scoring procedures when dealing with conjunctions, but not with disjunctions?

Although we take the results presented below to be relevant regardless of which answers one gives to these questions, one possible interpretation is suggested by the phenomenon of *homogeneity* observed for plurals in natural language.

Plural entities in natural language have the logically surprising feature that negation applies homogeneously to each individual in the plurality. Consider the examples of plurals in (1) and (2), and the putative interpretations for the negated plural (2) in (2a) and (2b).

(1) The boys did their homework.

(2) The boys didn't do their homework.

a. None of the boys did his homework.
b. At least one of the boys didn't do his homework.

Sentence (1) means that *every* boy did his homework, with some tolerance for exceptions which needn't concern us here (Križ & Spector, 2021). Sentence (2) then ought to be simply the negation of (1), which would amount to the interpretation paraphrased in (2b). Yet, the negated plural in (2) has a much stronger interpretation, to the effect paraphrased in (2a). In general, negated plurals are interpreted in this unexpected way, from the standpoint of classical logic (Krifka, 1996; Lappin, 1989; Löbner, 2000). This observation applies to plurals as in (2), generated by a noun phrase with plural morphology "the boys," but also to plurals formed by means of an explicit conjunction: a sentence like "John and Mary don't speak German" means that neither John nor Mary speak German, not merely that at least one of John or Mary doesn't speak German (but see Szabolcsi & Haddican, 2004, for evidence of cross-linguistic variation on the available interpretations).

In light of these observations, we hypothesize that participants in our experiment might score candidate explanations for losses by tracking the contributions of candidate causes to the stronger losing conditions obtained by negating the sufficient winning conditions in equation 1, which is a disjunction of plural terms, in this non-standard way. This would amount to the stronger loss conditions at the end of equation 3 below, which we preface with '$\not\equiv$' to indicate that it violates classical-logical equivalence.

$$\begin{aligned} \text{LOSS} &:= \neg((A \wedge B) \vee (C \wedge D)) \\ &\equiv \neg(A \wedge B) \wedge \neg(C \wedge D) \\ &\not\equiv \neg A \wedge \neg B \wedge \neg C \wedge \neg D = \text{LOSS}_{\text{strong}} \end{aligned} \tag{2}$$

To be very clear, we are *not* saying that we expect people will *misinterpret* the rules of the game. We do not expect that they would classify, say, a round of the game where the player only draws colored balls from urns A and C as anything other than a loss, any more than we expect English speakers that take (2a) to be the natural interpretation for (2) would mistake (1) for a true sentence, had exactly one boy done his homework. Instead, our hypothesis is that the tendency to negate plurals in this homogeneous way shapes the model that participants use as target to match for computing the contribution of various causes to losses across counterfactuals. Just like people assess a cause's contribution to wins by matching it against sufficient conditions $\{A \wedge B,\ C \wedge D\}$, they score its contribution to losses by matching it against the plural negation of those, which yields LOSS_{strong}.

Methods

Design and materials

The methodology was similar to that of Experiment 1. We presented participants with a simple game of chance. This time, the game involved four urns, with two different colors, purple and yellow (Figure 5). We randomized the assignment of colors, but always in such a way that urns A and B were of one color, and urns C and D of the other color. To win a round of the game, one needed to draw "two purple balls or two yellow balls."

While we randomized the specific urns' indices and their spatial arrangement for each participant, for simplicity here we refer to a consistent arrangement as depicted in Figure 5, where urn A has 14 colored balls, urn B 2, urn C 4, and urn D 19 colored balls. These induce different prior probabilities of drawing a colored ball out of each urn, such that $P(A) = 0.7$, $P(B) = 0.1$, $P(C) = 0.2$, and $P(D) = 0.9$. Throughout the experiment, the urn containing 14 colored balls and the urn containing 2 colored balls were always of the same color, while the other two urns (19 and 4 colored balls) were of the other color, so that each color would contain one high probability and one low-probability urn.

Procedure and participants

As in Experiment 1, participants first had the opportunity to familiarize themselves with the game and the rule determining a winning outcome, as well as with the underlying probabilities, by playing the game for ten rounds, as in Figure 5a. Urn draws and outcomes at this stage were pseudo-randomized in such a way as to reflect the underlying probabilities.

After they played ten rounds of the game, they saw the outcomes of rounds played by another player named John (as in Figure 5b) and were asked to rate on a Likert scale from 1 to 9 the causal responsibility of certain events, both singular and plural. Specifically, we queried their causal judgments by asking them the extent to which they agreed (on a 1–9 scale) with a sentence that followed the template: "John won (/lost) because he drew colored (/white) balls from box(es) [XYZ]." Figure 5c shows an example.

All participants saw four different rounds of the game played by John, one at a time, and provided their judgments after each round. All the rounds were played with the same underlying rule and the same urns in the same display as the one participants had been familiarized with. Each trial differed only in the outcome of the draw made by John. We presented all participants with the following four rounds, in random order:

1. Overdetermined positive: John drew a colored ball from each of the four urns — John won (Figure 5b)

2. triple-positive: John drew a colored ball from urns A, B, and D, but not from urn C — John won (Figure 5d)

3. triple-negative: John drew a white ball from urns A, B, and D, but not from urn C — John lost (Figure 5e)

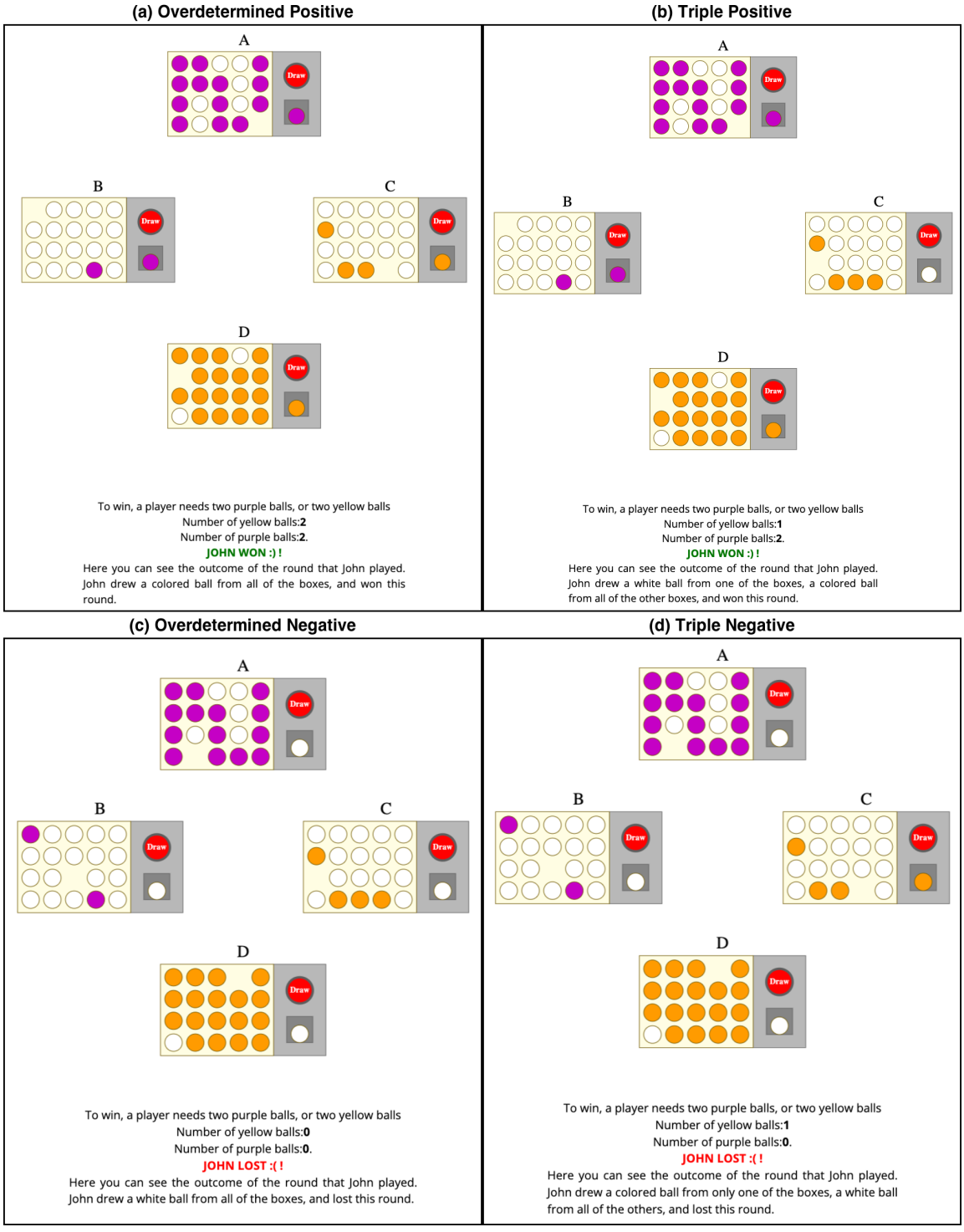
4. Overdetermined negative: John drew a white ball from all four urns — John lost (Figure 5f).

Within each round, we asked participants about every singular event that featured a colored ball in the winning rounds, and every singular event that featured a white ball in the losing rounds. We also asked about every plural combination of these singulars, with the exception of four-variable plurals (we considered those questionable candidates for causal selection judgments, since they provided an exhaustive description of all drawing events in a given round) and other plurals which we considered redundant with some that we already asked. The questions were presented in random order, with no separation between singulars and plurals.

We recruited a total of 368 participants (153 male, 215 female, mean age: 37.3) from all English-speaking countries on Prolific. We excluded from analysis 57 participants who failed to correctly answer either one of our two elementary comprehension questions, yielding

Figure 5

The four different outcomes presented to participants in Experiment 2. The familiarization phase, where people would get used to the underlying rule of the game and to the probabilities, was the same for all rounds. The questions were asked in a format analogous to that of Experiment 1.



a final sample of 311 participants whose data we analyzed. Each participant answered all of the questions of the four conditions in this experiment.

Computational modeling

We computed the predictions of the CESM and the NSM following the same procedure as in Experiment 1. We fitted the value of s and γ for both models by finding the parameter values that resulted in the best fit between model judgments and average participant judgments across all four conditions. As in Experiment 1, we used a grid search, exploring a wide range of values for the parameter s, crossed with different values for a scaling parameter γ. For the CESM, the best-fitting value was $s = 0.21$ (with $\gamma = 0.39$). For the NSM we find $s = 0.02$ (with $\gamma = 0.28$).

We also explored a variant of the computational models that allows for the possibility that participants handle the losing cases in the non-classical fashion discussed above. We provide the details of this model in the relevant subsection of the results.

Results

We first go through the results for each round separately. We start each section by a brief exposition of the predictive performance of the CESM and NSM models for the round, before delving into a qualitative analysis of the relevant patterns of judgments observed for that round. Note that none of the patterns we identify or the interpretation we provide for them depend on the models considered, unless explicitly specified otherwise. We provide these predictions mainly for readers interested in how state-of-the-art counterfactual models fare at predicting these new data.

Winning rounds

Overdetermined positive round. In this round, the player drew a colored ball from each of the four urns (as in Figure 5b) and therefore won the game.

Figure 6 summarizes our results in this condition. The CESM had a moderate but positive fit to participants' average judgments, $r(8) = 0.45$, while the NSM predictions were uncorrelated with participants' judgments, $r(8) = -0.18$.

Participants' judgments also reveal the following patterns.

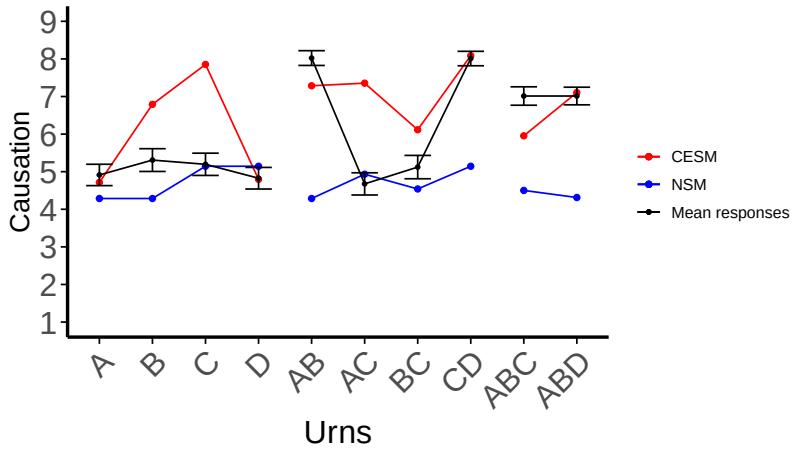
Non-linearity. Participants judged that B and C were the most important singular causes. Therefore, a linear combination approach would predict that they should also view the pair $B \wedge C$ as the best plural cause. In fact, participants judged that the pairs $A \wedge B$ and $C \wedge D$ were significantly better causes than $B \wedge C$, in clear opposition to the predictions of the linear combination hypothesis.

Participants preferred non-crossing over crossing pairs. There was a clear preference for pairs that did not cross the disjunction ($A \wedge B$, $C \wedge D$) over those that featured one variable on each side of the disjunction (e.g. $A \wedge C$, $B \wedge C$), (mean non-crossing: 7.02, mean crossing: 4.90; $t(1055.63) = 23.97$, $p < 0.0001$).

Weak abnormal inflation for singular variables. We observed an abnormal-inflation effect at the level of singulars, meaning that participants deemed urns B and C, which contained the lowest proportion of colored balls, more important for bringing about the outcome. Formally, judgments for B and C were higher than for A and D, $t(1239.25) = -2.56$,

Figure 6

Participants' responses, along with model predictions, for the Overdetermined positive *round. The red line represents the CESM predictions, the blue line the NSM predictions, and the black line represents the mean of participants' responses.*



$p = 0.010$. This qualitative pattern aligned with the predictions of the CESM, but not with the predictions of the NSM, which prescribed abnormal deflation in this context. No significant difference could be observed however between the two low-probability singulars, contrary to the CESM's expectations (means: 5.31, 5.19; $t(619.67) = 0.52$, $p > 0.6$).

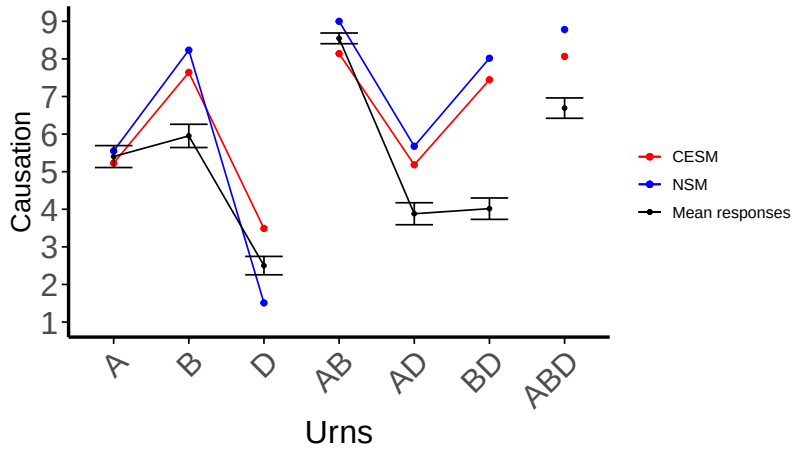
The CESM overestimates the attractiveness of some plurals. The CESM mistakenly predicted that $A \wedge C$ should be rated higher than $A \wedge B$, and $A \wedge B \wedge D$ higher that $A \wedge B \wedge C$. In both cases, the predictions come from a tendency of the model to give a very similar rating to the singular X and the pair $X \wedge Y$ if Y is a high-probability variable. This is because if $P(Y)$ is high, the correlation between $X \wedge Y$ and the outcome is very similar to the correlation between X and the outcome. This property often results in erroneous predictions, not only in this particular round, but also in the Triple-positive round below, where plurals containing the variable D are overestimated. We come back to this pattern in the Discussion section for this experiment.

Triple-positive round. In this round, the player drew a colored ball from urns A, B and D (as in Figure 5d) and therefore won the game. In such a draw, the win is not overdetermined like it was in the previous round, but clearly it is caused by the player's getting a colored ball from both *purple* urns A and B. Urn D, on the other hand, is not an active cause of the win in the present world, because it has no effect on winning in the absence of C.

Notice that, in the particular context of this round, drawing a colored ball from urn D does not simply have a low impact on the win, but in a categorical sense it is not at all a cause of the outcome in the actual world. A standard view of how causal selection judgments work holds that only the events that can be counted as *actual causes* (Halpern, 2016) of the outcome qualify as candidates for causal selection in the first place (see for example Gerstenberg et al., 2021; Quillien & Lucas, 2023). Following this logic, the causal impact score of the event "drawing a colored ball from urn D" should simply be zero, and it is unclear if plural events that contain D (such as "drawing colored balls from urns A and D") should

Figure 7

Participants' responses, along with model predictions, for the triple-positive *round. The red line represents the CESM predictions, the blue line the NSM predictions, and the black line represents the mean of participants' responses.*



count as actual causes or not. For simplicity, we gloss over this issue, allowing the model to give non-zero causal responsibility to D or plurals that feature D.

Figure 7 summarizes the results for the triple-positive rounds. Both counterfactual models give a good account of participants' judgments: model predictions are correlated with average human judgments $r(5) = 0.78$ (CESM) and $r(5) = 0.79$ (NSM). We now highlight the most significant patterns.

Abnormal inflation effect for singulars. We did observe an abnormal inflation effect, with the low-probability urn B being ranked significantly higher than high-probability urn A ($t(617.88) = -2.54$, $p = 0.011$), in line with the predictions of both the CESM and the NSM.

Ceiling-high ratings for the pair $A \wedge B$. Participants were almost unanimous in giving ceiling-high ratings to $A \wedge B$. Only 51 participants (out of 311) in total gave it ratings different from the maximal value of the Likert scale.

Low ratings for D, and plurals containing D. Ratings for the idle variable D were very low. More than half of participants (171 out of 311) gave it maximally low ratings. Interestingly however, the ratings weren't as low as they were high for $A \wedge B$, suggesting that the fact that D does make a contribution to the win in other possible configurations still had some residual influence on participants' ratings.

Plurals containing D, such as the mixed pairs $A \wedge D$, $B \wedge D$, and the triple $A \wedge B \wedge D$, were systematically rated somewhere between the best cause that they contained and the low ratings of D. They were systematically rated lower than predicted by the models, which didn't penalize strongly enough the inclusion of the idle variable D. However, participants didn't seem to systematically disqualify a plural just for including the variable D (for example, by giving it ratings as low as those of D alone).

Losing rounds

The first two conditions just discussed collected judgments about the contribution of *colored ball* draws to a player's *win* in a given round of the game. The two conditions we present next instead queried participants' judgments on the contribution of *white ball* draws to a player's *loss*.

Earlier we discussed the hypothesis that when participants make judgments about a loss, they might simulate counterfactual possibilities in a slightly different way than when they judge a win. In order to formalize this hypothesis in a counterfactual framework, we consider a variant of our computational models featuring a parameter w, which encodes participants' propensity to represent the losing conditions in the non-classical, language-like way depicted in equation 3, reproduced below for convenience.

$$\text{LOSS}_{\text{strong}} := \neg A \wedge \neg B \wedge \neg C \wedge \neg D \tag{3}$$

By contrast, the equation for losing obtained by applying the classical negation of the equation for winning is as follows.

$$\begin{aligned} \text{LOSS} &:= \neg((A \wedge B) \vee (C \wedge D)) \\ &\equiv \neg(A \wedge B) \wedge \neg(C \wedge D) \end{aligned} \tag{4}$$

Concretely, we assume that when the outcome under consideration is a loss, the participant makes a random decision in each counterfactual world, where:

- with probability w, the loss is determined non-classically (equation 3);
- otherwise, with probability $1 - w$, the loss is determined by the classical negation of the original rule (equation 4). This entails that our earlier w-less models can be understood as a special case of the w models where $w = 0$;
- once it has been determined whether a given world is an instance of a win or a loss, the worlds that are not losses are recorded as wins. The impact of each variable on the models is then computed exactly as before.

We fitted the models again in this new version using data from all four conditions, via a three-dimensional grid search (s, w, γ). The best fitting values were respectively $s = 0.21$ and $w = 0.77$ (with $\gamma = 0.41$) for the CESM, and $s = 0.5$ and $w = 0.77$ (with $\gamma = 1.17$) for the NSM. For simplicity, all model predictions we report use the values of the s and γ parameter fitted jointly with w, even for the base versions. Using the original fitted parameters for the base versions yields virtually identical results.

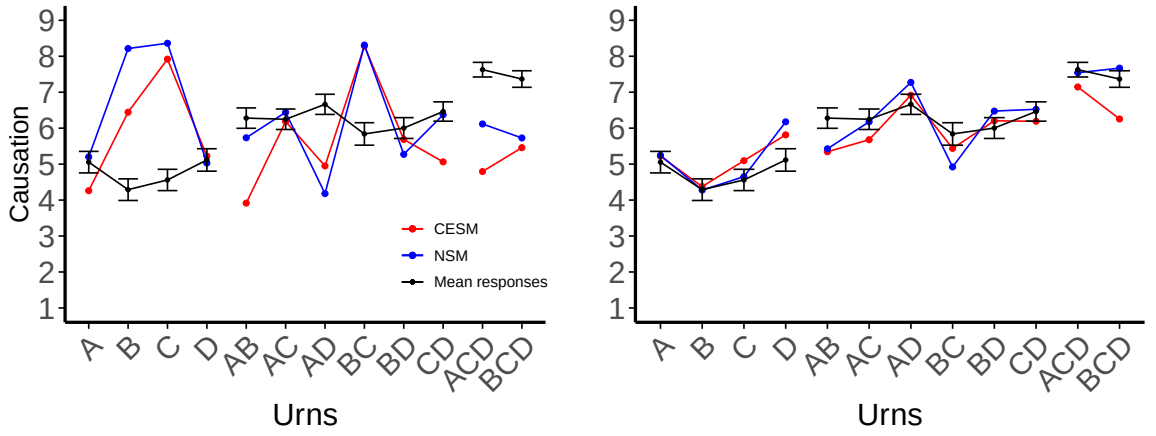
Adding the w parameter significantly improved the fit of both models, even accounting for differences in degrees of freedom (Table 6). For the negative conditions below, we report both versions of the models, to showcase the impact of the new parameter.

Overdetermined negative condition. The Overdetermined negative condition is the mirror image of the Overdetermined positive condition. Here, the player drew a white ball from all four urns, and consequently lost, as pictured in Figure 5e.

The results are summarized in Figures 8a and 8b. The base versions of the CESM and NSM have a poor fit to participants' average judgments, $r(10) = -0.38$ (CESM) and $r(10) = -0.42$ (NSM). In contrast, the versions of the models featuring the w parameter provide a good account of the data, $r(10) = 0.82$ (CESM) and $r(10) = 0.87$ (NSM); see also Table 7. We now go over our most telling findings.

Figure 8

Participants' responses and model predictions for the Overdetermined negative *round. The red line represents the CESM predictions, the blue line the NSM predictions, and the black line represents the mean of participants' responses.*



(a) *Base models without w parameter*

(b) *Models with w parameter, encoding participants' tendency to represent the losing conditions non-classically*

Urns with the lowest number of white balls are given higher scores. This effect can be observed both for singulars and for plural causes, with combinations featuring urns A or D scoring higher than those featuring B or C. This pattern runs completely contrary to the predictions of counterfactual models under the classical representation of losing conditions from equation 4, but is captured by the version that assumes a non-classical representation of the losing conditions.

Indeed, if participants are representing losing conditions as a disjunction of minimally sufficient conditions (as in equation 4), we would expect their judgments to follow the logic of abnormal *deflation* and ascribe a greater causal impact to those urns out of which one is most likely to get a white ball, that is urns B and C. Instead, their judgments seem to follow a logic of abnormal *inflation*, with a preference for the urns that contain the lowest number of white balls, i.e. A, D, consistent with a representation of the losing conditions as a conjunction of necessary events as in equation 3.

No significant difference between pairs that cross the disjunction and those that do not. Participants' judgments for pairs that crossed the disjunction (e.g. A and C) were not significantly different than for pairs that did not cross the disjunction (mean crossing: 6.19; mean noncrossing: 6.37; $t(1311.86) = -1.47$, $p = 0.140$).

This finding is consistent with the idea that participants represent the losing conditions as $\text{LOSS}_{strong} := \neg A \wedge \neg B \wedge \neg C \wedge \neg D$, with no natural grouping of the variables. In contrast, a classical representation of the losing conditions would have predicted that any pair of events on the same side of the purple vs. yellow divide should be redundant, since a single white ball on either side is sufficient to cancel any contribution that this side could have made to a win. There is no such redundancy however if the representation is non-classical, where each

white-ball drawing event makes a crucial contribution to the outcome.

Triples are rated higher than pairs. Mean pairs: 6.25; Mean triples: 7.37; $t(1400.78) = -12.64$, $p < 0.0001$. Here again, while triples would have been redundant under a classical representation, each element of the triple makes a non-zero contribution to the outcome if the representation is non-classical.

Triple-negative condition. In the triple-negative round, the player drew white balls from every urn except for urn C, as in Figure 5f. This makes it a mirror image of the Triple-positive round, where white balls are substituted for colored balls. In this round, the white ball from urn D is indispensable for the loss, whereas urns A and B are redundant with one another.

The same contrast between classical and non-classical representations of losing conditions applies in this round. Here, the w parameter that we enriched our models with encodes participants' propensity to represent the rule as follows.

$$\text{LOSS} := \neg A \wedge \neg B \wedge \neg D$$

We take it that the non-classical representation of the losing conditions in this round is slightly different from the Overdetermined negative round because, in the actual world, a colored ball *was* drawn from urn C. This makes the negation of the plural entity $C \wedge D$ in our rule harder to represent as the strong plural negation $\neg C \wedge \neg D$, since the situation at hand is known to be one where the player in fact drew a colored ball from urn C. In other words, the player cannot possibly have lost *because* they drew a *white* ball from C, since they in fact drew a *colored* ball from C.

Results are summarized in Figures 9a and 9b, and in Table 7. The base versions of the CESM and NSM have a moderate fit to participants' average judgments, $r(10) = 0.34$ (CESM) and $r(10) = 0.52$ (NSM). On the other hand, the versions of the models featuring the w parameter provide a good account of the data, $r(10) = 0.94$ (CESM) and $r(10) = 0.99$ (NSM); see also Table 7. We highlight some of the most important qualitative patterns below.

Participants prefer urns with a lower number of white balls. Causal judgments for $\neg A$ were higher than $\neg B$ ($t(618.98) = 2.98$, $p = 0.003$), and causal judgments for $\neg A \wedge \neg D$ were higher than $\neg B \wedge \neg D$ ($t(619.49) = 2.34$, $p = 0.020$). The preference for urns featuring a lower number of white balls is similar to what we find in the Overdetermined negative round. Again this pattern is most coherent with a non-classical representation of the losing conditions.

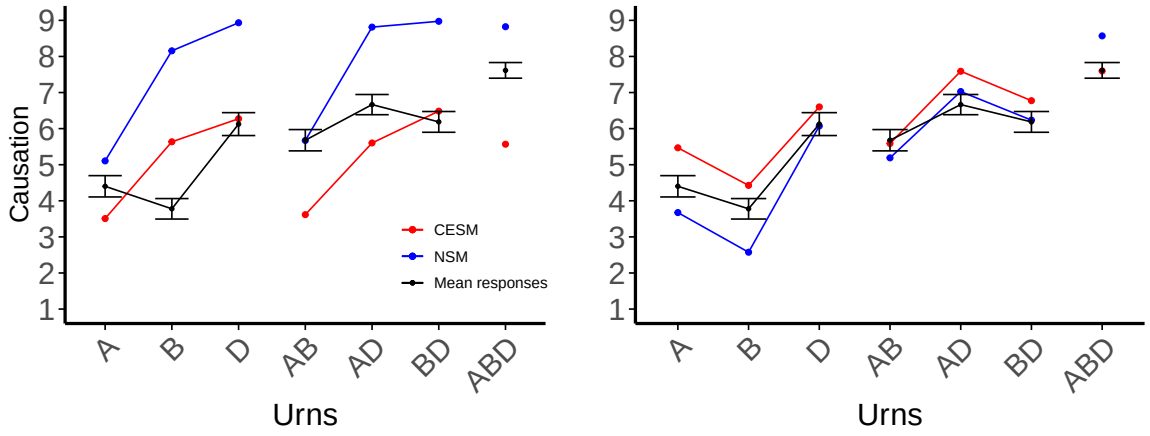
The pair $\neg A \wedge \neg B$ rates higher than either of its constitutive singulars, and the triple $\neg A \wedge \neg B \wedge \neg D$ rates higher than its constitutive pairs ($t(708.60) = 10.27$, $p < 0.0001$). Both patterns are examples of plurals whose effect in the outcome under the classical representation is redundant with that of one of the events (singular or plural) contained within it, which should lead them to be rated at most as high as the sufficient event in question. The fact that these are rated higher by participants is again suggestive of their representing the losing conditions non-classically.

Overall model comparison

Table 6 summarizes the comparison between the models at the global level (all conditions combined). The version of the CESM that includes the w parameter has the best fit overall (BIC = 53835.36; correlation with means: $r(35) = 0.67$, $p < 0.001$). This is

Figure 9

Participants' responses and model predictions for the triple-negative *round. The red line represents the CESM predictions, the blue line the NSM predictions, and the black line represents the mean of participants' responses.*



(a) *Base models without w parameter* (b) *Models with w parameter*

Table 6

Model comparisons for Experiment 2, across all conditions. The Cor. *column indicates the item-level correlation between model predictions and mean participant responses per question.*

Model	LogLik	χ^2	*p*-value	BIC	Cor.
Baseline	-27494			54997.32	0
CESM, with *w*	-26905	1175.91	< 0.0001 ***	53840.57	0.67
CESM, no *w*	-27963	2023.44	< 0.0001 ***	55889.97	0.2609
NSM, with *w*	-27634	642.55	< 0.0001 ***	55295.31	0.57
NSM, no *w*	-28390	11511.80	< 0.0001 ***	56967.07	0.02

better than the fit of the model without the w parameter (BIC = 55926.62; correlation with means: $r(35) = 0.26$, $p < 0.001$), or than any of the versions of the NSM model (with w: BIC = 55295.31, cor.: $r(35) = 0.57$, $p < 0.001$; without w: BIC = 56967.07, cor.: $r(35) = 0.02$). In general, the versions of the models that include the w parameter are better than the versions without it, by all metrics.

We also compared these models with a constant baseline model, which always made the same predictions for every question in every condition. The prediction was fitted to the data via the scaling parameter γ only. All counterfactual models had a better fit than the baseline model when assessed in terms of their correlations with mean human judgments, but only the version of the CESM that included the w parameter had a better BIC score than the baseline model.

Table 7

Model fits per condition. The Cor. *column indicates the item-level correlation between model predictions and mean participant responses per question.*

Condition	Model	BIC	AIC	Cor.
Overdetermined positive	CESM	15009.14	14991.01	0.45
	NSM	15278.45	15260.32	-0.18
Triple-positive	CESM	10335.21	10318.16	0.78
	NSM	10505.71	10488.66	0.79
Overdetermined negative	CESM, no w	19238.14	19225.69	-0.38
	CESM, w	17731.07	17718.62	0.82
	NSM, no w	19164.41	19151.96	-0.42
	NSM, w	17703.51	17684.84	0.87
triple-negative	CESM, no w	10853.57	10842.19	0.34
	CESM, w	10335.21	10318.16	0.94
	NSM, no w	10655.7	10644.33	0.52
	NSM, w	10422.88	10405.82	0.99

Discussion

This second experiment provides more evidence in favor of the psychological reality of plural causes in the context of causal selection judgments.

Just like in the first experiment, participants' judgments for plural causes across all four rounds of the game were clearly sensitive to the probabilities attached to the corresponding events. Participants' judgments in the Overdetermined positive round corroborate the non-linearity between participants' judgments for plurals and their judgments for the singular causes that constitute them. Given the pattern of abnormal inflation observed for singular variables, favoring B and C over A and D, a *linear* reconstruction of participants' judgments for plurals would have us expect the pair $B \wedge C$ to rank above all others pairs, when in fact it ranks much lower than the $A \wedge B$ and $C \wedge D$ pairs.

The winning rounds of the experiment also demonstrate that plural causes featuring more variables are *not necessarily* rated higher than proper subsets of the variables they contain. The Triple-positive round shows a clear pattern in this regard: every time a plural features the variable D, its rating is systematically lower than that of the same cause (singular or plural), minus the variable D. This contradicts the hypothesis that adding more variables always makes an explanation more attractive, which the results from Experiment 1 could not rule out. And the phenomenon is not limited to the situation where an idle variable like D features in a plural: a similar observation can be made about the triplets $A \wedge B \wedge C$ and $A \wedge B \wedge D$ in the overdetermined positive condition, both of which are rated lower than the best pair that they contain, $A \wedge B$. Thus, although plurals featuring more variables might be descriptively more thorough, they can still be unappealing if their overall counterfactual dependence profile drops as a result of the variables added.

We also uncovered properties of plural causal judgments that go beyond what is ex-

pected based purely on patterns of counterfactual dependence. First, in the winning rounds of our second experiment, participants dislike causal explanations that "cross" the disjunction $(A \wedge B) \vee (C \wedge D)$, above and beyond what is predicted by counterfactual models. The fact that the causal rule features two clearly distinct sufficient conditions seems to exert an influence on participants' explanatory preferences not fully captured by the counterfactual dependence profile of the variables in question.

Second, the following property of the CESM was not reflected in participants' judgments. The model tends to give a very similar rating to a singular X and the pair $X \wedge Y$ if Y is a high-probability variable. This is because if $P(Y)$ is high, the correlation between $X \wedge Y$ and the outcome is very similar to the correlation between X and the outcome. This property often results in erroneous predictions, like in the Overdetermined positive round where the model predicts (against participants' judgments) that the pair $A \wedge C$ should rate higher than the pair $B \wedge C$.

Finally, we found that counterfactual models could only account for participants' judgments in the losing rounds if we assume that participants handle the losing conditions, in their internal computations, in a way inconsistent with the classical-logical negation of the winning conditions. Specifically, participants seem to be representing the negation of the winning conditions (i.e. the losing conditions) in a way consonant with how natural-language represents the negations of plurals.

General discussion

Humans make systematic judgments regarding which of several events influencing an outcome should be considered as *the cause*, or the most important cause of that outcome. These *causal selection* judgments are the object of a rich and actively expanding section of the psychological literature on actual causation. So far, however, this literature has been exclusively focused on *singular* events, identified with the distinct nodes of the relevant causal system. In this article, we argue that its scope should be extended to include *plural* events, featuring multiple variables.

Our experiments present strong evidence that judgments about plural events cannot be captured in terms of linear combinations of the judgments for the events that constitute them. There appears to be no obvious way of combining participants' causal judgments regarding any two events A and B that would predict their judgment for the event "*A and B*." Our results thus establish the psychological reality of plural causes: plural causes are treated by the mind as causal entities in their own right, and their impact on the outcome is apprehended in a *holistic* or wholesale fashion. The lesson here is that people's assessment (and likely also their production) of causal explanations is not constrained by the boundaries of individual variables but instead can operate at the level of multivariate groupings.

A second set of observations looks at the ways in which such groupings differ from classical conjunctions of variables. In positive outcomes, groups of variables that cross over sufficient conditions for producing an outcome seem to be disfavored. Even more striking are judgments in negative outcomes, where the default expectations of counterfactual models are reversed, unless one assumes that subjects assess causes by their contribution to a representation of losses that is altogether baffling under standard classical-logical expectations, but understandable if plural causes function like plurals in natural language.

Summary of our findings and their immediate consequences

Plural cause judgments cannot be reconstructed as linear combinations of singular judgments

It seems *prima facie* plausible that, when people make a causal judgment about whether "A and B caused E," they might judge how much A caused E, judge how much B caused E, and then combine these two judgments into a single judgment for the plurality. Under this view, plural causal selection would be entirely predictable from facts about singular causal selection. One of our main goals was to rule out this null hypothesis.

In our two experiments we designed situations in which computational models predict that, if plurals are processed in a holistic manner, judgments about plural causes should not be simple combinations of judgments about their constituent singular variables. Participants' judgments in these situations supported this prediction, successfully ruling out the linear combination hypothesis.

Counterfactual models can account for a broad range of plural causation judgments

A growing body of research provides strong evidence that causal judgment involves counterfactual thinking (e.g. Gerstenberg et al., 2017; Kahneman & Miller, 1986; Krasich et al., 2024; Quillien & Lucas, 2023). At the same time, there are debates about what phenomena counterfactual theories can explain (Hall, 2004; Henne, 2023; Lombrozo, 2010; Rose et al., 2021; Sytsma, 2020), and about the computations that counterfactuals might be an input to (Icard et al., 2017; Quillien, 2020).

Our experiments provide a rich opportunity to probe the scope and the generality of counterfactual theories. None of the counterfactual theories that we are aware of were developed with the goal of explaining data about how people make plural causation judgments. Consequently, accounting for these judgments off-the-shelf would constitute important evidence in favor of these theories.

We found that two recent counterfactual models of causal judgment (Icard et al., 2017 and in particular Quillien and Lucas, 2023) can quantitatively account for many features of participants' judgments. In particular, when participants' judgments about plurals diverge from a linear combination of their constituent singular causes, they typically do so in the way that is predicted by the counterfactual models. As such, our results strengthen the case for counterfactual theories.

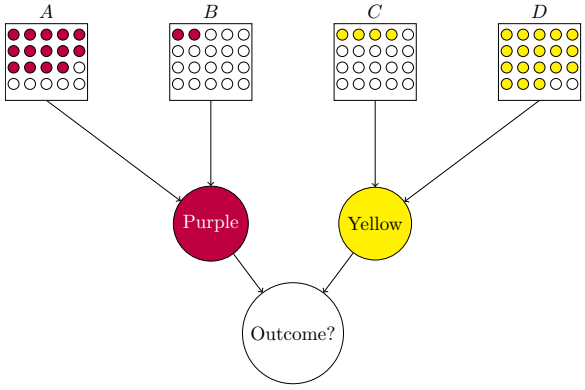
The losing rounds of our second experiment did however offer a more ambivalent verdict. On the one hand, counterfactual models used "out of the box" performed very poorly and even predicted patterns of preferences that were the exact opposite of those observed. On the other hand, the same models could be made to track those patterns very well when complemented with a specific assumption about how people score causes in negative outcomes, more on which shortly. The plausibility of a counterfactual account of these data thus depends on the plausibility of this additional assumption.

Causal judgments favor sets of variables belonging to the same disjunct

In the game that participants played in Experiment 2, the player needed to draw either two purple or two yellow balls in order to win the game. From a logical point of view, this

Figure 10

Causal graph representing participants' putative model of the situation.



rule is a *disjunction*: the player wins if either one of the conditions for victory is met; each condition is a *disjunct*.

Participants favored plural causes that do not cross the boundary between the two disjuncts. Suppose for example that the player drew purple balls from urns A and B and yellow balls from urns C and D. In this situation, participants would be reluctant to say that the player "won because he drew a colored ball from urn B and urn C." The counterfactual models also disfavored these boundary-crossing plurals, but participants did so to an even greater extent than predicted.

There are different possible explanations for this pattern. At a superficial level, for example, participants might have preferred causal explanations that mentioned balls of the same color because of low-level perceptual biases.

At a deeper level, participants might have built an internal representation of the game in terms of a causal model with a particular structure. This causal model would contain intermediate variables (in the technical, causal-model sense of "variable," Pearl, 2000; Woodward, 2016) representing whether each condition for victory (getting two purple balls and getting two yellow balls) is met, see Figure 10. Such a model would be distinct from one without the intermediate variables, in that it would support new kinds of interventions unavailable to the more straightforward model. For example, one could in principle intervene to set the variable corresponding to "Both yellow balls," i.e. $A \wedge B$, to TRUE even though one yellow ball is not present. But this seems like a paradoxical causal model, as it features variables that hold a containment relation to one another (see e.g. Beckers et al., 2023, for how this can introduce problems for structural models in general).

An alternative, related hypothesis is that people represent the groupings $A \wedge B$, $C \wedge D$ not as distinct intermediate causal variables, but as distinct processing *pathways* for deriving outcomes from variables A, B, C, D in accordance to the causal model WIN $:= (A \wedge B) \vee (C \wedge D)$. This hypothesis is in line with accounts in the mental model theory tradition (see in particular Khemlani et al., 2018) which propose that people represent disjunctions by representing each disjunct within a 'mental model'. For example, to represent the disjunction $(A \wedge B) \vee (C \wedge D)$, people would picture two distinct mental models, $A \wedge B$ and $C \wedge D$, and assess whether the disjunction is true by checking whether the conditions specified by one of the model are

satisfied.

There is independent evidence that this structured way of representing disjuncts can account for several surprising patterns of reasoning with disjunctions (Chung et al., 2022; Koralus & Mascarenhas, 2013; Walsh & Johnson-Laird, 2004). Studies of deductive reasoning have investigated how people reason about logical statements of the shape $(A \wedge B) \vee C$. In experiments replicated and varied multiple times, participants overwhelmingly conclude B from the two premises $(A \wedge B) \vee C$, and A (Koralus & Mascarenhas, 2018; Picat & Mascarenhas, 2024; Sablé-Meyer & Mascarenhas, 2021; Walsh & Johnson-Laird, 2004). But this is a fallacy: it is compatible with the premises but not the conclusion that A and C should be true while B is false. Koralus and Mascarenhas (2013) explain this fact in terms of question-answer dynamics: the disjunction in the first premise is naturally interpreted as demanding the participant *choose* between one of the two disjuncts. This in turn induces dependencies between propositions occurring *within disjuncts*: in the context of $(A \wedge B) \vee C$, the second premise A is seen as an answer in the $A \wedge B$ direction, introducing dependence between A and B. In general, this approach predicts that the conjuncts within each disjunct will be taken, as it were, to *hang together* in a cohesive way, so that learning about one will constitute evidence in favor of all of the others.

There is even evidence of such effects absent the language of disjunction, in experiments where the same information was conveyed by means of visual stimuli in the form of animations (Chung et al., 2022), indicating that this "packaged" way of representing a disjunction is not simply a fact about the interpretation of the word "or" and its equivalent locutions. Rather, these rich, structured disjunctive representations which induce dependencies not predicted by standard Boolean interpretations of logical connectives are available to human minds far more generally. In particular, they might have been available to participants in our Experiment 2, and may have played a part in shaping their causal judgments, by pushing them to associate A and B on the one hand and C and D on the other more tightly than is predicted by classical accounts of disjunction, whether deductive or probabilistic.

Plural causes featuring more events are not necessarily better

In Experiment 1, we found that participants preferred causal explanations that mentioned the most causes, and that this preference was stronger than predicted by counterfactual models. We probed the extent of this trend in Experiment 2, where we found that mentioning more causes does not always make a causal explanation better. For example, a causal explanation mentioning only two events A and B might be judged better than an explanation mentioning A, B, and C.

These results suggest that causal judgments are subject to a trade-off between two different considerations (see also Kirfel et al., 2024; Sumers et al., 2024). On the one hand, people might favor explanations that give detailed information about what events happened. Since every explanation of the shape "X happened because Y" comes with the implication that "Y happened" (Halpern, 2016), causal explanations that feature many causes offer more complete descriptions of what happened. With respect to this criterion, plural cause explanations are always more helpful than singular ones, since they highlight more true facts about the situation.

On the other hand, causal explanations convey information about patterns of counter-

factual dependence (Quillien, 2020). Under this criterion, large plural causes can sometimes be worse. For example, an explanation mentioning three events A, B, and C might misleadingly suggest that the outcome strongly covaries with the conjunction of these three events, across counterfactuals.

Judgments for the losing rounds suggest a non-standard representation of losing conditions

In the trials of Experiment 2 where the player loses the game, we could not account for the data by assuming that participants internally represent the conditions for losing the game (for the purpose of simulating counterfactual possibilities) as the classical complement of the conditions for winning. Instead, judgments can be captured quite adequately if we suppose that, in the losing rounds, participants score various causes to the extent that they contribute to a different target, which is a situation in which the player draws a white ball from all urns.

This target does not correspond to the classical-logic negation of the winning conditions and does not track with what people know to be the actual minimal conditions for losing a round of the game. But it is consistent with the assumption that, to assess causes of negative outcomes, people track how various causes match with the *plural* negation of the disjunction or set of models $\{A \wedge B, C \wedge D\}$ they would use to assess their contribution to positive outcomes. This suggests that the representation participants use for deriving an outcome might possess a semantics similar to that of natural-language plurals, in that these conjunctions are treated as homogeneous entities, and negated as such. Note, in this connection, that the effects we find are unlikely to stem from linguistic experimenter demands when interpreting the causal statements verbally. We asked participants about the causal impact of "drawing a white ball" on a "loss," never about the impact of "not drawing a colored ball" on "not winning."

Additional theoretical and experimental work is needed to put this particular hypothesis about plural negation on firmer ground, and our general thesis about the psychological reality of plural causes does not depend on it. For now, we observe that this hypothesis is a first step toward incorporating into models of causal judgment more sophisticated considerations about the representational arsenal at humans' disposal when simulating counterfactuals. Typically, counterfactual models have assumed that we can gloss over the particular way that humans represent logical connectives ("and," "or," and so on) when they simulate counterfactual possibilities, with the implicit premise that these representations jibe well enough with the normative standards from classical logic. In contrast, our results suggest that the details of the language of thought (Fodor, 1975; Quilty-Dunn et al., 2023) and attending *logic* of thought might make a big difference for our theories, and involve the kinds of non-trivial, non-classical properties that are found in natural language. More generally, our account of people's judgments in the losing rounds illustrates the fact that work on the formal semantics of natural language is a fruitful source of hypotheses about psychological phenomena (Wellwood & Hunter, 2023).

Conclusion

The current study is a preliminary demonstration that, to the mind, causes are more than the sum of their parts. Judgments about plural causes are affected by the prior probability of their constituent variables, but cannot be derived from the causal strength of these individual

variables. Our results are consistent with simple extensions of extant counterfactual models of causal selection. At the same time, our findings raise new issues about the psychology of causation, which point to opportunities for extending existing theories. We anticipate that future research on the psychology of plural causation will open yet further avenues of inquiry about human causal cognition.

References

Arendt, H. (1987). Collective responsibility. In J. W. S. Bernauer (Ed.), *Amor mundi: Explorations in the faith and thought of Hannah Arendt* (pp. 43–50). Springer, Dordrecht. https://doi.org/10.1007/978-94-009-3565-5_3

Beckers, S., Halpern, J. Y., & Hitchcock, C. (2023, April). Causal models with constraints. In M. van der Schaar, C. Zhang, & D. Janzing (Eds.), *Proceedings of the second conference on causal learning and reasoning* (pp. 866–879, Vol. 213). PMLR. https://proceedings.mlr.press/v213/beckers23a.html

Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*, *37*(6), 1171–1191. https://doi.org/10.1111/cogs.12062

Chung, W., Bade, N., Blanc-Cuenca, S., & Mascarenhas, S. (2022). Question-answer dynamics in deductive fallacies without language. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th annual meeting of the cognitive science society*. https://escholarship.org/uc/item/9711612q

De Leeuw, J. R. (2015). Jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Fodor, J. (1975). *The language of thought*. Harvard University Press.

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*. https://doi.org/10.1037/rev0000281

Gerstenberg, T., & Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599.

Gerstenberg, T., Lagnado, D., & Zultan, R. (2023). Making a positive difference: Criticality in groups. *Cognition*, *238*, 105499. https://doi.org/10.1016/j.cognition.2023.105499

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

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Gill, M., Kominsky, J. F., Icard, T. F., & Knobe, J. (2022). An interaction effect of norm violations on causal judgment. *Cognition*, *228*, 105183.

Griffiths, T., & Tenenbaum, J. (2005). Structure and strength in causal induction. *Cognitive psychology*, *51*, 334–84. https://doi.org/10.1016/j.cogpsych.2005.05.004

Hall, N. (2004). Two concepts of causation.

Halpern, J. Y. (2015). A modification of the Halpern-Pearl definition of causality. *arXiv preprint*. https://doi.org/10.48550/arXiv.1505.00162

Halpern, J. Y. (2016). *Actual causality*. MIT Press.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. part i: Causes. *British Journal for the Philosophy of Science*, *56*(4), 843–887. https://doi.org/10.1093/bjps/axi147

Henne, P. (2023). Experimental metaphysics: Causation. *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*, 104708.

Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019). A counterfactual explanation for the action effect in causal judgment. *Cognition*, *190*, 157–164.

Hesslow, G. (1988). The problem of causal selection. In *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 11–32).

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, *98*, 273–299. https://doi.org/10.2307/2678432

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*(5), 942–951. Retrieved April 15, 2024, from http://www.jstor.org/stable/10.1086/667899

Icard, T. F. (2016). Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, *7*(4), 863–903. https://doi.org/10.1007/s13164-015-0283-y

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93. https://doi.org/10.1016/j.cognition.2017.01.010

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

Khemlani, S., Byrne, R. M. J., & Johnson-Laird, P. N. (2018). Facts and possibilities: A model-based theory of sentential reasoning. *Cognitive Science*, *42*(6), 1887–1924. https://doi.org/10.1111/cogs.12634

Kinney, D. B., & Lombrozo, T. (2024). Building compressed causal models of the world. *Cognitive Psychology*. https://osf.io/preprints/psyarxiv/2f7x6

Kirfel, L., Harding, J., Shin, J. Y., Xin, C., Icard, T., & Gerstenberg, T. (2024). Do as i explain: Explanations communicate optimal interventions. *Proceedings of the annual meeting of the cognitive science society*, *46*.

Kirfel, L., Icard, T., & Gerstenberg, T. (2021). Inference from explanation. *Journal of Experimental Psychology: General*, *151*. https://doi.org/10.1037/xge0001151

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology*. MIT Press.

Kominsky, J., & 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.

Kominsky, J., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209. https://doi.org/10.1016/j.cognition.2015.01.013

Koralus, P., & Mascarenhas, S. (2013). The erotetic theory of reasoning: Bridges between formal semantics and the psychology of deductive inference. *Philosophical Perspectives*, *27*, 312–365. https://doi.org/10.1111/phpe.12029

Koralus, P., & Mascarenhas, S. (2018). Illusory inferences in a question-based theory of reasoning. In K. Turner & L. Horn (Eds.), *Pragmatics, truth, and underspecification:*

Towards an atlas of meaning (pp. 300–322, Vol. 34). Leiden: Brill. https://doi.org/10.1163/9789004365445_011

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.

Krifka, M. (1996). Pragmatic strengthening in plural predications and donkey sentences. In T. Galloway & J. Spence (Eds.), *Proceedings of SALT 6* (pp. 136–153). https://doi.org/10.3765/salt.v6i0.2769

Križ, M., & Spector, B. (2021). Interpreting plural predication: Homogeneity and non-maximality. *Linguistics and Philosophy*, *44*, 1131–1178. https://doi.org/10.1007/s10988-020-09311-w

Lappin, S. (1989). Donkey pronouns unbound. *Theoretical Linguistics*, *15*(3), 263–289. https://doi.org/10.1515/thli.1988.15.3.263

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

Löbner, S. (2000). Polarity in natural language: Predication, quantification and negation in particular and characterizing sentences. *Linguistics and Philosophy*, *23*, 213–308. https://doi.org/10.1023/A:1005571202592

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257. https://doi.org/10.1016/j.cogpsych.2006.09.006

Lombrozo, T. (2010). Causal-explanatory pluralism: How intensions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 302–332. https://doi.org/10.1016/j.cogpsych.2010.05.002

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299. https://doi.org/10.1016/j.cognition.2013.12.010

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

Miller, S. (2001). Collective responsibility. *Public Affairs Quaterly*, *15*(1), 65–82. https://www.jstor.org/stable/40441276

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PLoS ONE*, *14*(8). https://doi.org/10.1371/journal.pone.0219704

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, *111*(2), 455–485. https://doi.org/10.1037/0033-295X.111.2.455

O'Neill, K., Henne, P., Quillien, T., Icard, T., & DeBrigard, F. (2025). Norms moderate causal judgments in cases of double prevention. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

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

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, *146*(12), 1761–1780. https://doi.org/10.1037/xge0000318

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books, Inc. https://doi.org/10.5555/3238230

Picat, L., & Mascarenhas, S. (2024). On the interplay between interpretation and reasoning in compelling fallacies. *Cognitive Science*, *48*(12). https://doi.org/10.1111/cogs.70021

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).

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The re-emergence of the Language of Thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, *46*(e292). https://doi.org/10.1017/S0140525X22002849

R Core Team. (2013). *R: A language and environment for statistical computing* [ISBN 3-900051-07-0]. R Foundation for Statistical Computing. Vienna, Austria.

Rose, D., Sievers, E., & Nichols, S. (2021). Cause and burn. *Cognition*.

Sablé-Meyer, M., & Mascarenhas, S. (2021). Indirect illusory inferences from disjunction: A new bridge between deductive inference and representativeness. *Review of Philosophy and Psychology*, *12*(2). https://doi.org/10.1007/s13164-021-00543-8

Sloman, S. A., & Lagnado, D. A. (2015). Causality in thought. *Annual Review of Psychology*, *66*, 223–247.

Sumers, T. R., Ho, M. K., Griffiths, T. L., & Hawkins, R. D. (2024). Reconciling truthfulness and relevance as epistemic and decision-theoretic utility. *Psychological review*, *131*(1), 194.

Sytsma, J. (2020). Causation, responsibility, and typicality. *Review of Philosophy and Psychology*, *12*(4), 699–719. https://doi.org/10.1007/s13164-020-00498-2

Szabolcsi, A., & Haddican, B. (2004). Conjunction meets negation: A study in cross-linguistic variation. *Journal of Semantics*, *21*(3), 219–249. https://doi.org/10.1093/jos/21.3.219

Walsh, C., & Johnson-Laird, P. N. (2004). Coreference and reasoning. *Memory and Cognition*, *32*(1), 96–106. https://doi.org/10.3758/BF03195823

Wellwood, A., & Hunter, T. (2023). Linguistic meanings in mind. *Behavioral and Brain Sciences*, *46*(e289). https://doi.org/10.1017/S0140525X23001887

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*. https://doi.org/10.1215/00318108-115-1-1

Woodward, J. (2016). The problem of variable choice. *Synthese*, *193*(4), 1047–1072.