Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 8, 2142–2159
https://doi.org/10.1037/xge0001615

Modeling Confidence in Causal Judgments

Kevin O'Neill[1, 2], Paul Henne[3, 4], John Pearson[1, 2, 5, 6], and Felipe De Brigard[1, 2, 7]

[1] Center for Cognitive Neuroscience, Duke University
[2] Department of Psychology and Neuroscience, Duke University
[3] Department of Philosophy, Lake Forest College
[4] Neuroscience Program, Lake Forest College
[5] Department of Biostatistics and Bioinformatics, Duke University
[6] Department of Electrical and Computer Engineering, Duke University
[7] Department of Philosophy, Duke University

Counterfactual theories propose that people's capacity for causal judgment depends on their ability to consider alternative possibilities: The lightning strike caused the forest fire because had it not struck, the forest fire would not have ensued. To accommodate a variety of psychological effects on causal judgment, a range of recent accounts have proposed that people probabilistically sample counterfactual alternatives from which they compute a graded measure of causal strength. While such models successfully describe the influence of the statistical normality (i.e., the base rate) of the candidate and alternate causes on causal judgments, we show that they make further untested predictions about how normality influences people's *confidence* in their causal judgments. In a large (N = 3,020) sample of participants in a causal judgment task, we found that normality indeed influences people's confidence in their causal judgments and that these influences were predicted by a counterfactual sampling model in which people are more confident in a causal relationship when the effect of the cause is less variable among imagined counterfactual possibilities.

Public Significance Statement

People are thought to identify an event as a cause of an effect when altering it would make a difference to the effect. Despite stable patterns in causal judgments across scenarios, however, people often disagree about the causes of particular effects. Here, we asked how people determine their confidence in such judgments, and we found evidence that people are more confident in their judgments when the difference made by the cause to the effect is robust to changes in background factors.

Keywords: causal judgment, counterfactual thinking, metacognition, confidence

Supplemental materials: https://doi.org/10.1037/xge0001615.supp

Judgments about cause and effect are central to the way people decide who or what is responsible for an outcome (Chockler & Halpern, 2004; Malle et al., 2014; Sytsma, 2021, 2022) and explain how a particular state of affairs came to be (Lombrozo, 2007; Lombrozo & Vasilyeva, 2017). Psychologists have found that causal judgments are affected by normality (i.e., the extent to which an event conforms to statistical, social, or moral norms; Gerstenberg & Icard, 2020; Henne, O'Neill, et al., 2021; Icard et al., 2017; Knobe & Fraser, 2008), the presence of alternative causes (Kominsky et al., 2015; Lagnado et al., 2013; O'Neill, Henne, et al., 2022), temporal recency (Bramley et al., 2018; Henne, Kulesza, et al., 2021; Spellman, 1997), action–omission differences (Henne et al., 2019),

Kevin O'Neill https://orcid.org/0000-0001-7401-9802
Paul Henne https://orcid.org/0000-0002-3526-2911
John Pearson https://orcid.org/0000-0002-9876-7837
Felipe De Brigard https://orcid.org/0000-0003-0169-1360

This research was supported by the Office of Naval Research (Grant N00014-17-1-2603) awarded to Felipe De Brigard. The authors also thank Benjamin Eva and Elika Bergelson for their helpful comments on various versions of this article. The authors have no conflicts of interest to declare.

Data, code, and materials are available at https://osf.io/sm6qg/. Previous versions of this article were presented at the Neural Information Processing Systems Workshop on Metacognition in the Age of AI: Challenges and Opportunities, the Cognitive Science Society, the Society for Philosophy and Psychology, and the Southern Society for Philosophy and Psychology.

Kevin O'Neill played a lead role in conceptualization, data curation, formal analysis, methodology, software, visualization, writing–original draft, and writing–review and editing. Paul Henne played a supporting role in conceptualization, methodology, supervision, validation, and writing–review and editing. John Pearson played a supporting role in conceptualization, formal analysis, methodology, supervision, validation, visualization, and writing–review and editing. Felipe De Brigard played a lead role in funding acquisition and supervision and a supporting role in conceptualization, investigation, resources, validation, and writing–review and editing.

Correspondence concerning this article should be addressed to Kevin O'Neill, Department of Psychology and Neuroscience, Duke University, 417 Chapel Drive, Reuben-Cooke Building, Durham, NC 27708, United States. Email: kevin.oneill@duke.edu

and foreseeability (Kirfel & Lagnado, 2021). Taken together, these findings suggest that while certain factors affect people's judgments, people generally agree about causal relationships in particular scenarios. In the current article, we focus on the influence of normality on causal judgments, where people tend to judge an event as more causal when it violates a statistical or moral norm (Kominsky & Phillips, 2020).

People, however, often disagree about the extent to which events should be judged as causal, even when they are confident in their judgments (Gerstenberg et al., 2021; O'Neill, Henne, et al., 2022). For instance, economists and politicians often argue about whether a recent rise in inflation is due to financial policies or rather to supply chain issues, with each side confident in their judgment. Such disagreements are especially critical to resolve because decision making is not only informed by causal judgments (Hagmayer & Sloman, 2009; Hitchcock & Knobe, 2009; Morris et al., 2018) but also by confidence (Dotan et al., 2018; Folke et al., 2016; Yeung & Summerfield, 2012). That is, without an immediate way to determine which side is right and which side has misplaced their confidence, it is unclear what to do to prevent future inflation. So, the existence of such disagreements leads to a central, though largely unstudied, question: How do people evaluate the accuracy and reliability of their own causal judgments about single events?

Here, we propose a possible answer to this question: Confidence in causal judgments indicates the robustness of the counterfactual relationship between a candidate cause and an effect. That is, people should be confident that an event caused an effect when it always makes a difference to the effect, and they should be confident that it did not cause the effect when it never makes a difference to the effect. Conversely, people should be uncertain whether an event caused an effect when it only makes a difference to the effect in the presence of a number of background conditions. Our account predicts that causal judgments and confidence should be nonlinearly related because confidence should be high for causal judgments that are very low (indicating the absence of a causal relation) or high (indicating a strong causal relation) but low for causal judgments that are in between (indicating a weak causal relation). By relying on this predicted nonlinear relationship between confidence and causal judgments, we also show that confidence can help arbitrate between alternative mechanisms of causal judgment.

To answer these questions, we will first review counterfactual sampling models of causal judgment. We will then draw on Bayesian models of metacognition in perception and decision making (Fleming & Daw, 2017; Ma & Jazayeri, 2014; Meyniel & Dehaene, 2017; Meyniel et al., 2015; Pouget et al., 2016) to endow counterfactual sampling models with a normative metric of confidence in causal judgments. Following the predictions of our extension of counterfactual sampling models, we then tested whether participants' confidence ratings were sensitive to statistical normality and causal structure. Along with replicating previously found effects on the mean and variability of causal judgments, we found that confidence ratings exhibited qualitatively similar effects and that these effects were simultaneously predicted by one of the models we reviewed (i.e., the Necessity–Sufficiency model; Icard et al., 2017). Finally, we argue in the discussion that this constitutes strong evidence in favor of counterfactual sampling models, and we discuss the implications of this work for future research on causal judgments and metacognition.

Counterfactuals and Causal Judgment

Imagine that Joe is playing a simple game where he randomly selects balls from two boxes. The left box contains 30% green and 70% red balls, and the right box contains 60% blue and 40% orange balls. The rules of the game are simple: If Joe selects both a green ball from the left box and a blue ball from the right box, he wins a dollar. Joe simultaneously chose a green and blue ball from the two boxes, and so he won a dollar. To what degree did Joe win the dollar because he picked a green ball (Morris et al., 2019)?

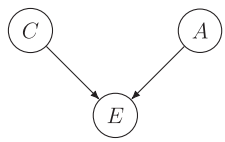
Here, we focus on counterfactual accounts of causal judgment because they offer a precise answer to this question. Inspired by metaphysical (Hume, 1748; Lewis, 1974) and statistical (Pearl, 2009) theories of causation, counterfactual accounts of causal judgment assume that people represent general causal dependencies between types of events using a causal graph. Figure 1 depicts the causal graph for the game above, which is known as an unshielded collider (Pearl, 2019). In this graph, an effect E (winning a dollar) is produced by two causes: a candidate cause C (picking a green ball from the left box) and an alternate cause A (picking a blue ball from the right box). Here, we will focus on two causal structures over this graph. In the *conjunctive* structure, both C and A need to occur for the effect E to occur (as in the example above). In the *disjunctive* structure, either C or A alone (or both C and A) is sufficient for the effect E. Here we call C the candidate cause because it is the event that we aim to judge.

Given a causal graph, counterfactual accounts assume that people make causal judgments by evaluating whether the effect would have been different under varying circumstances. In the example above, Joe would not have won the dollar had he not picked a green ball. So, picking a green ball made a difference to whether he won the dollar, and one can say it caused him to win the dollar (Hart & Honoré, 1985; Lewis, 1974). But many other combinations of events could have happened in principle. If he did not pick a blue ball, for example, Joe would not have won a dollar no matter whether he picked a green ball or not. In this case, picking a green ball would not have made a difference to whether he won the dollar because the outcome would be the same either way. Owing to this feature of counterfactual thinking, theorists have noted that counterfactual accounts need a way to evaluate the difference made by the candidate cause in a wide range of possible alternatives (Hitchcock, 2012; Lombrozo, 2010; Quillien, 2020).

To account for the fact that the effect of the candidate cause can depend on the values of other variables, recent accounts (which we refer to as *counterfactual sampling models*; Figure 2A, gray boxes) assume an iterative form of the same underlying logic known as *Monte Carlo sampling*. First, people use the causal structure, the prior probabilities of different events, and information about what

Figure 1

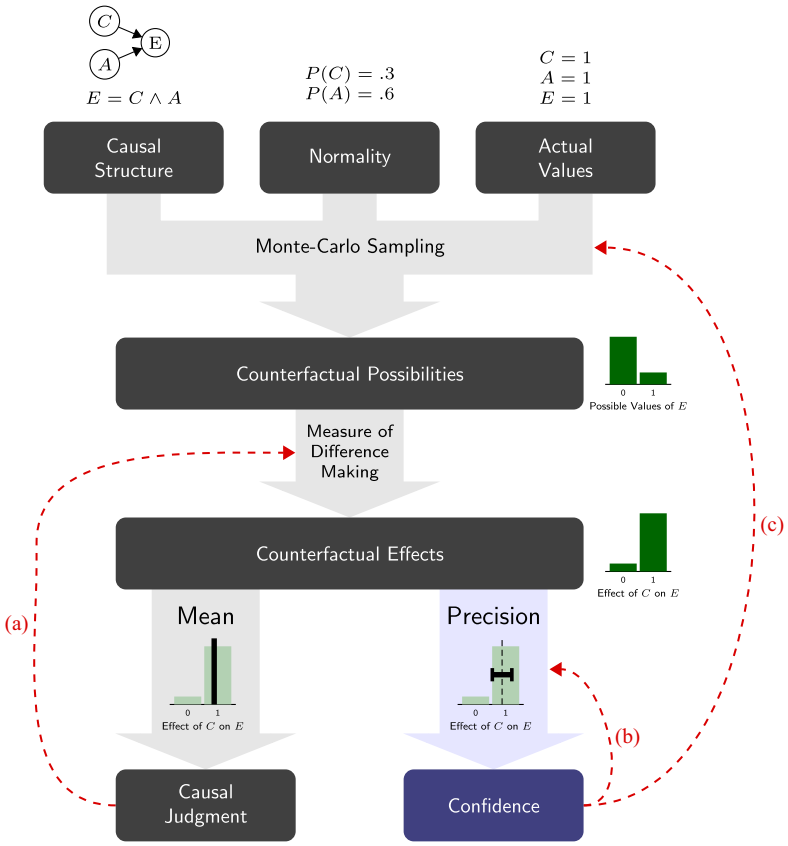
A Causal Graph Depicting the Relationships Between an E as Produced by a C and an A



Note. E = effect; C = candidate cause; A = alternate cause.

Figure 2

The Counterfactual Sampling Model of Causal Judgments (Gray) and Our Extension of This Model to Confidence (Blue)



Note. Boxes indicate relevant constructs and box arrows indicate assumed relationships between constructs. Given information about the causal structure, the probabilities of different events, and what actually happened, people imagine a distribution of counterfactual possibilities and then determine the difference the C made to the E across each possibility. People report causal judgments as the mean counterfactual effect, and they report their confidence as the precision (inverse variance). Discrepancies between predicted and observed judgments can be resolved by amending one of the above assumptions (revisions depicted as red dashed arrows). When a model fails to predict causal judgments, theorists usually amend the measure of difference making (a). When a model fails to predict confidence, theorists have two main alternatives. First, they could argue that even if causal judgments are normative, confidence is determined nonnormatively (b). Second, they could argue that confidence is normative, but people generate counterfactual possibilities according to some alternative sampling mechanism (c). E = effect; C = candidate cause; A = alternate cause. See the online article for the color version of this figure.

actually happened to imagine many different counterfactuals in proportion to their perceived likelihood (Kahneman & Miller, 1986). For each imagined possibility, they determine whether the effect happens and whether intervening on the candidate cause would have changed the effect, providing a distribution of causal effects specific to that counterfactual. Finally, these counterfactual effects are averaged to form a causal judgment (Gerstenberg et al., 2021; Icard et al., 2017; Quillien, 2020). This process of generating and averaging counterfactuals provides a Monte Carlo estimate of the subjective probability that the candidate cause made a difference to the effect (Icard, 2016). Counterfactual sampling models thus have the desirable property that, given an appropriate way of quantifying difference making, the very same mechanism can be used to predict a wide variety of patterns in causal judgments (Henne, 2023).

Recent work has focused on using discrepancies between predicted and observed causal judgments to identify which kind of difference making best explains causal judgments (depicted as dashed arrow a in Figure 2A; Morris et al., 2019; O'Neill, Quillien, & Henne, 2022; Quillien, 2020). However, there remains disagreement on this matter because several different models tend to make qualitatively similar predictions of causal judgments.

So, rather than focusing on a particular formulation of counterfactual sampling models, here we will provide a model of confidence applicable to the entire class of recent models and some classic models of causal judgment that can be reasonably construed under this framework (Cheng, 1997; Cheng & Novick, 1990; Spellman, 1997). In the following section, we first review five successful measures of difference making, which can be subsumed under counterfactual sampling models (Table 1).

Measures of Difference Making

Here, we will outline five ways of quantifying difference making proposed in psychology and philosophy that we selected based on their ability to produce quantitative predictions of the effects of normality on causal judgments: the ΔP (Cheng & Novick, 1990; Jenkins & Ward, 1965), Power PC (the causal power theory of the probabilistic contrast model; Cheng, 1997), Crediting Causality (Spellman, 1997), Necessity–Sufficiency (Icard et al., 2017), and Counterfactual Effect Size models (Quillien, 2020). Table 1 depicts the formula for each model along with the derived model predictions for the conjunctive and disjunctive causal structures. Notably, each is expressed as a probability measure using an intervention on a causal graph (i.e., the *do* operator; Pearl, 2009). Mathematically, the intervention $do(C = c)$ involves fixing C to a particular value c and removing the influence of any causes of C; here we use the shorthand $do(C)$ when fixing C to 1 and $do(\neg C)$ when fixing C to 0. Interventions are generally not equivalent to simple conditioning, for example, $P(E|C)$, because conditioning on C merely selects cases where C has a certain value without removing the influence of other variables on C. In the unshielded collider, however, C has no causes (i.e., it has no incoming edges), so conditioning and interventions are equivalent in this context. Nevertheless, we will use the intervention notation as it is more general.

The simplest measure, known as ΔP, is the average difference in the effect when the candidate cause is introduced, which is denoted as $P(E|do(C))$, compared with when it is removed, which is denoted as $P(E|do(\neg C))$ (Cheng & Novick, 1990; Jenkins & Ward, 1965). In our example, Joe wins a dollar if he gets both a green ball and a blue ball. Given that he gets a green ball, the probability that he wins a dollar is just the probability that he gets a blue ball. Without the green ball, this probability is 0 (he needs both balls to win the dollar). So, ΔP predicts that drawing a green ball makes a difference to Joe winning the dollar when he draws a blue ball (i.e., $\Delta P = P(A)$).[1]

The Power PC model extends ΔP with the intuition that the candidate cause can only be said to generate the effect in the subset of cases in which the effect would not already be generated by some other causes. Accordingly, it is equal to ΔP normalized with respect to the probability that the effect does not occur when the candidate cause is removed, that is, $P(\neg E|do(\neg C))$ (Cheng, 1997). In our example, the Power PC model makes the same predictions as ΔP (i.e., that drawing a green ball makes a difference when Joe also draws a blue ball) because Joe never wins the dollar without drawing a blue ball.[2]

The Crediting Causality model (Spellman, 1997) is also similar to the ΔP model, but it uses the unconditional probability of the effect overall as a baseline. So, it can be interpreted as the increase in the probability of the effect when the cause is present compared with the probability of the effect in general, allowing it to account for some, but not all, effects of temporal recency on causal judgments (Henne, Kulesza, et al., 2021). In our example, the overall probability that Joe wins a dollar is the product of the probability that he draws a green ball and the probability that he draws a blue ball. Subtracting this product from the probability of winning a dollar given that Joe draws a green ball, we see that the Crediting Causality model predicts that drawing the green ball makes a difference when he draws a blue ball, but not a green ball, that is, $CC = P(A)(1 - P(C))$. Like ΔP and the Power PC model, the Crediting Causality model was not derived with a commitment that these probabilities are estimated through counterfactual sampling. However, because counterfactual sampling is a plausible mechanism by which they can be estimated, we include the Crediting Causality model here.

The Necessity–Sufficiency model computes the impact of the candidate cause by taking a weighted average of the degree to which it is necessary and sufficient for the effect (Icard et al., 2017). Specifically, it predicts that when people imagine the cause to occur, they compute sufficiency by checking whether the effect occurs. When they imagine the cause as absent, they compute necessity by checking whether the effect does not occur. In our example, drawing a green ball is sufficient for Joe to win a dollar only when he also draws a blue ball. Drawing a green ball is also completely necessary

[1] Note that the ΔP model was originally formulated to handle judgments of general causation (e.g., judgments of the probability that a medicine prevents a symptom within a population), whereas here we are interested in causal judgments of singular events (e.g., judgments that a medicine prevented a particular individual's symptom). As a result, it was not derived with counterfactual sampling in mind because participants typically were presented with many individual cases over which to generalize and, therefore, did not need to imagine alternative possibilities (Cheng & Novick, 1990; Jenkins & Ward, 1965). But since it directly corresponds to the average treatment effect (i.e., regression coefficient between the candidate cause and the effect; Pearl, 2009) and since many other models are straightforward modifications of it (e.g., Cheng, 1997; Quillien, 2020; Spellman, 1997), ΔP provides a reasonable quantity people could be estimating when using counterfactual sampling in making causal judgments.

[2] The formula provided in Table 1 assumes that the causes of an effect occur independently and independently influence the effect. Notably, the assumption of independent influence is not met in our conjunctive example above, which requires both the green ball and the blue ball to win a dollar. Novick and Cheng (2004) introduced a variant of the Power PC model that allows for conjunctive causes, but this model would predict that the green ball alone is never a cause of Joe winning a dollar because only the conjunction of the green ball and the blue ball is causal. Likewise, as with ΔP, the Power PC model was originally derived as a measure of general causal power, not to model the causal judgments of singular events in which we are interested. It was later extended to handle judgments of single events (Cheng & Novick, 2005; Stephan et al., 2020; Stephan & Waldmann, 2018), though in our example these versions of the model predict that picking a green ball is always maximally causal. Finally, several Bayesian extensions of the Power PC model have been proposed which incorporate uncertainty about the causal graph (Griffiths & Tenenbaum, 2005; Holyoak et al., 2010; Lu et al., 2008; Stephan et al., 2020; Stephan & Waldmann, 2018; Tenenbaum & Griffiths, 2001). As we will be focusing on cases in which the causal graph is fully known to the participant, these extensions reduce to the standard Power PC model. Thus, since these later versions of the model either reduce to the original Power PC model or are unable to predict any of the established normality effects in which we are interested, here we use the original formulation of the Power PC model provided by Cheng (1997). As in Morris et al. (2019), our goal in including this model is not to claim that the overall epistemic status of the model (as originally formulated) hinges on its predictions of singular causal judgments; rather it is simply to determine whether the measure of difference making suggested by the model in the context of general causal judgments can provide similarly useful predictions in the context of counterfactual sampling models of singular causal judgments.

Table 1
Causal Strength Metrics for the Unshielded Collider With Generative Causes (Figure 1) From Five Counterfactual Sampling Models

Model	Measure of difference making	Conjunctive	Disjunctive
ΔP (Jenkins & Ward, 1965)	$P(E\|do(C)) - P(E\|do(\neg C))$	$P(A)$	$1 - P(A)$
Power PC (Cheng, 1997)	$\Delta P / P(\neg E\|do(\neg C))$	$P(A)$	1
Crediting Causality (Spellman, 1997)	$P(E\|do(C)) - P(E)$	$P(\neg C)P(A)$	$P(\neg C)P(\neg A)$
Necessity–Sufficiency (Icard et al., 2017)	$P(C)P(E\|do(C)) + P(\neg C)P(\neg E \mid do(\neg C, A))$	$P(C)P(A) + P(\neg C)$	$P(C)$
Counterfactual Effect Size (Quillien, 2020)	$\Delta P \sigma_C / \sigma_E$	$\sqrt{\frac{P(\neg C)P(A)}{1-P(C)P(A)}}$	$\sqrt{\frac{P(C)P(\neg A)}{P(C)+P(A)-P(C)P(A)}}$

Note. E = effect; C = candidate cause; A = alternate cause; PC = probabilistic contrast; σ_C = standard deviation of C; σ_E = standard deviation of E.

for him to get the dollar. So, overall, the Necessity–Sufficiency model says that Joe drawing a green ball makes a difference to him winning a dollar when he draws both a green and a blue ball or when he does not draw a green ball: In other words, it makes a difference except in the case where he draws a green ball, but not a blue ball. Though it was developed specifically to account for normality effects on causal judgment (e.g., Henne, 2023; Kominsky & Phillips, 2020), the Necessity–Sufficiency model has also been shown to account for interactions of normality effects (Gill et al., 2022), temporal recency effects (Henne, Kulesza, et al., 2021), action–omission effects (Henne et al., 2019), as well as judgments in more complex causal structures including double prevention (Henne & O'Neill, 2022).

Finally, in our causal structure of interest, the Counterfactual Effect Size model measures difference making as ΔP standardized with respect to the standard deviations of the candidate cause and effect, σ_C and σ_E (Quillien, 2020). Just as ΔP can be interpreted as the regression coefficient between the candidate cause and the effect across the considered counterfactual possibilities, the Counterfactual Effect Size model can be interpreted simply as the correlation between the cause and the effect across the considered possibilities. So, the Counterfactual Effect Size model has the convenient interpretation that when people make causal judgments, they are computing the effect size of the cause within the imagined possibilities. While it is relatively new, the Counterfactual Effect Size model has been shown to outperform the above models in predicting normality effects on causal judgment, as well as causal judgments about elections (O'Neill, Quillien, & Henne, 2022; Quillien, 2020; Quillien & Barlev, 2021).

Counterfactual Sampling and Metacognition

If people make causal judgments by (a) considering counterfactual possibilities through Monte Carlo sampling, (b) computing counterfactual effects by determining whether the cause made a difference to the effect in each counterfactual, and (c) averaging together these counterfactual effects, two questions naturally arise with respect to confidence: First, *why* might people keep track of confidence in their causal judgments? Second, *how* might they estimate confidence in such judgments? In this section, we motivate and introduce a model of confidence in causal judgments by answering these two questions.

The Role of Confidence in Causal Judgment

Many tasks in cognitive science ask participants to make a decision after receiving some evidence. For instance, a researcher might ask a participant to decide whether a stimulus is currently present or not, whether they have previously seen a stimulus or not, or whether to accept option A or B. A key insight of recent work on metacognition is that in such tasks, people's decision confidence generally tracks the probability that their decision was correct (Fleming & Daw, 2017; Hangya et al., 2016; Kepecs & Mainen, 2012; Kiani & Shadlen, 2009; Peters, 2022; Pouget et al., 2016), though this correspondence is not exact (Peters et al., 2017; Samaha & Denison, 2020; Shekhar & Rahnev, 2021). Moreover, researchers have shown that momentary decreases in confidence also predict information-seeking behavior (Desender et al., 2018; Goupil et al., 2016), error-monitoring processes (Boldt & Yeung, 2015; Yeung & Summerfield, 2012), and subsequent changes in mind (De Martino et al., 2013; Resulaj et al., 2009). Finally, some argue that metacognition plays an important role in social coordination and group decision making (Heyes et al., 2020; Pescetelli et al., 2016).

If people rely on internal estimates of confidence to predict the accuracy of their decisions and the decisions of others, it is likely that confidence should play a similar role in the domain of singular causal judgment. There are, however, two key differences between singular causal judgments and typical tasks in metacognitive research that warrant explanation.

Perhaps the most apparent difference is that while many tasks involving metacognition have a clear criterion for the accuracy of a behavioral response, there is no such criterion in the domain of singular causal judgment. For instance, in a visual perception task, a participant's response of whether a stimulus was present or absent on a given trial is correct if and only if the stimulus *actually was* present on that trial. Thus, it makes sense that people would benefit from an estimate of the probability with which their answer was correct because they can use such an estimate to guide subsequent behavior whether or not they know the response is correct. Now consider the task of singular causal judgment: In the example above where Joe won a dollar after drawing a green and a blue ball, is the correct response that drawing the green ball caused Joe to win the dollar? Unlike the visual perception task, there is much reasonable disagreement about which answer is correct, and this disagreement is central to longstanding debates in the philosophy and psychology of causal judgment (Beebee et al., 2009; Godfrey-Smith, 2009; Henne & O'Neill, 2022; Lewis, 1974; Lombrozo, 2010; Wolff, 2007). Moreover, there is also debate as to whether there is a binary answer to this question or whether causal judgments are graded, further complicating this issue (Danks, 2017; Demirtas, 2022; Halpern & Hitchcock, 2015; Kaiserman, 2016, 2018; O'Neill, Henne, et al., 2022; Sartorio, 2020). Given that neither answer is

straightforwardly correct, it seems less plausible that people would require an estimate of the probability that their judgment is correct.

Second, while confidence often helps participants to calibrate learning across many related experiences, singular causal judgments (by definition) pertain to single events. In the visual perception task, for instance, a participant may integrate information about the presence or absence of a stimulus over many trials to form a belief about the overall prevalence of stimuli across the task and the extent to which they should update this belief on each trial depends on their current level of confidence. Theories of how people learn general causal relationships share this property: They assume that people have an internal model of whether and how much an effect generally depends on a cause and that people incrementally update this model in light of observed evidence according to Bayes' rule (Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001). Likewise, confidence in general causal judgments reflects how much participants' beliefs about causal relationships have updated in light of observed evidence (Liljeholm & Cheng, 2009; Perales & Shanks, 2003; Shanks, 1987). Critically, however, singular causal judgments are one shot. For instance, in studies of singular causal judgment, participants are given full knowledge of both the general causal structure and the sequence of events to be judged (Icard et al., 2017; Morris et al., 2019). This type of problem precludes learning across repeated observations because (a) all of the observable evidence is already known, and (b) repeated observations are impossible in practice (e.g., one cannot rewind time to recreate the exact conditions of a traffic accident). Without a need for learning across observations, it is unclear that people would require confidence to calibrate this kind of learning.

If confidence in singular causal judgments does not serve the role of providing an estimate of the probability of a correct judgment or of moderating learning across related experiences, it is unclear what role(s) it can still serve. To clarify this issue, we take inspiration from past research focusing on a kind of metacognition with both of these properties: metacognition about value-based judgments such as whether one would prefer to snack on an apple or an orange in the present moment (De Martino et al., 2013). Just like in causal judgments, there is no correct value-based judgment: People reasonably disagree about their momentary preferences. Value-based judgments are also singular: Given fixed and known general preferences, learning that someone chose an apple at one moment does not predict their choice at an unrelated time above and beyond these general preferences. Still, confidence in value-based judgments has been found to be related to choice consistency, making it important in predicting subsequent changes of mind (Folke et al., 2016).

By analogy, we argue that confidence in a singular causal judgment may be informed by the consistency of the effect of the cause across counterfactual possibilities—what is known in the literature as the *insensitivity* (Vasilyeva et al., 2018; Woodward, 2006), *robustness* (Gerstenberg et al., 2021; Grinfeld et al., 2020), or *portability* (Hitchcock, 2012; Lombrozo, 2010) of the causal relationship. In other words, people should be more confident in singular causal judgments when the candidate cause makes a similar difference to the effect across many imagined counterfactual possibilities compared with when the effect of the candidate cause on the effect varies widely across these possibilities. Consider again the introductory example. If it is guaranteed that Joe will draw a blue ball, drawing a green ball always allows him to win the dollar. So, in this case of a robust causal relationship, one can be confident that drawing the green ball caused him to win the dollar. But if Joe is equally likely to draw a blue ball or not—a less robust relationship—one should be less confident that drawing the green ball caused him to win the dollar because drawing the green ball only sometimes makes a difference. This framing helps explain why people might estimate confidence in singular causal judgments even when strict interpretations in terms of accuracy are not readily applicable to such estimates: Similar to confidence in value-based decisions, confidence in singular causal judgments carries information about how likely an identified causal relation is to generalize to novel circumstances. As a result, confidence in these judgments can still inform behavior like information seeking, changes of mind, social coordination, and group decision making (Peters, 2022).

A Precision Model of Confidence

If confidence in singular causal judgments is meant to signal the robustness of a causal relation across many different circumstances, how might people estimate confidence? The counterfactual sampling models of causal judgment introduced in the previous section assume that causal judgments are estimates of the degree to which a change in the candidate cause results in a change in the effect. Additionally, they make the further assumption that people compute these estimates as an average of a distribution of specific causal effects over a set of considered possibilities. But this distribution of specific causal effects can carry more information than just the average: In particular, it also carries information about its precision (i.e., its inverse variance). That is, if the considered possibilities concentrate tightly around some given value, the average provides a more or less complete description of the difference made by the candidate cause to the effect. If, in contrast, the candidate cause has a strong causal effect in some possibilities but a weak or even preventative effect in others, then the average provides a much less complete description of that effect. Overall, then, the precision of the distribution over specific causal effects provides a normative indicator of how well the average summarizes the entire distribution (Figure 2A, blue).

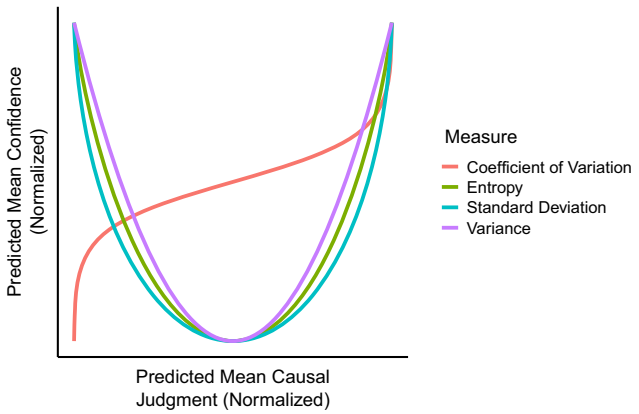
In the field of metacognition, confidence ratings of continuous decisions are often modeled in exactly this way (Liljeholm, 2015; Ma & Jazayeri, 2014; Meyniel & Dehaene, 2017; Meyniel et al., 2015; Navajas et al., 2017; Pouget et al., 2016; Yeung & Summerfield, 2012). Notably, this model of confidence assumes that people's confidence ratings are normative in the sense that they are higher for causal relationships that are robust than those that are not robust (Figure 2, dashed arrow b). This assumption may seem unnecessarily strong because it could be that confidence ratings instead reflect a simple heuristic or something else altogether. Counterfactual sampling models, however, already assume that people (a) imagine counterfactuals using Monte Carlo sampling and (b) compute a distribution of counterfactual effects when making judgments. It is difficult to see why people would undergo an intensive procedure to normatively produce causal judgments only to ignore this information when rating confidence. If the counterfactual sampling account is correct, then we can safely assume that confidence ratings reflect robustness. Overall, our proposal is that if people's causal judgments are an average of a distribution of causal effects specific to a set of considered possibilities, they should be

able to report their degree of confidence using the precision of this distribution.

To verify the robustness of our results and to highlight the flexibility of our account, we considered four different possible measures of precision: variance, standard deviation, entropy, and coefficient of variation.[3] In the causal structure of interest, the candidate cause either makes a difference to the effect or not within any single possibility. As a result, the counterfactual effects predicted by all models except the Counterfactual Effect Size model follow a Bernoulli distribution. Under the additional assumption that people sample approximately the same number of counterfactual possibilities across contexts, confidence can be calculated as a simple function of the mean (Figure 3; see Supplemental Table 1 for equations). The Counterfactual Effect Size model standardizes the specific causal effects by the standard deviation of the candidate cause and effect. So, for this model, confidence is also scaled by this constant. Overall, entropy, standard deviation, and variance suggest that confidence should be highest when the corresponding judgment is extreme (close to 0 or 1), whereas the coefficient of variation predicts that confidence should monotonically increase with causal judgments (Figure 3). While these relations hold in our causal structure of interest, we note that they only apply when the differences made by the candidate cause to the effect follow a Bernoulli distribution with a constant sample size: When the effect is a continuous variable, when the cause sometimes prevents the effect, or when people systematically imagine different numbers of counterfactuals across contexts, confidence may have a different relation to causal judgments.

Figure 3

The Predicted Relationship Between Mean Causal Judgment and Mean Confidence by Four Measures of the Precision of a Distribution in the Conjunctive/Disjunctive Unshielded Collider



Note. Because we allow a linear relationship between model predictions and empirical means, mean causal judgment and confidence are depicted on normalized scales. The entropy, standard deviation, and variance measures agree that confidence should be highest when the causal judgment is very high or very low and that confidence should be lowest when making intermediate causal judgments. In contrast, the coefficient of variation predicts that confidence should increase with causal judgments. For the Counterfactual Effect Size model, this relationship also depends on the estimated standard deviations of the candidate cause and effect. See the online article for the color version of this figure.

Understanding the Precision Model of Confidence

In this section, we demonstrate in a simple example how our precision model of confidence in causal judgments predicts related effects of normality (i.e., the statistical probability of an event) on causal judgments and confidence in those judgments. For simplicity, we focus here on the ΔP measure of difference making and the standard deviation measure of precision (other measures are estimated using the exact same process). Recall the earlier example where Joe wins a dollar if he draws a green ball from the left box and a blue ball from the right box, he draws both a green and a blue ball, and he wins the dollar (Figure 1). Did Joe win the dollar because he drew a green ball from the left box?

As discussed above, counterfactual sampling models predict that people answer this question by imagining a range of possibilities, where the probability of each event occurring in a given possibility is roughly proportional to its objective probability. Figure 4 depicts three such cases. In all three cases, there is a 50% likelihood that Joe will draw a green ball from the left box (the candidate cause, *C*). The likelihood that he will draw a blue ball from the right box (the alternate cause, *A*) is either 10% (Figure 4A), 50% (Figure 4B), or 90% (Figure 4C). Accordingly, counterfactual sampling models predict that among the alternative possibilities imagined when making a causal judgment, the probability of drawing a green ball is close to 50%, and the probability of drawing a blue ball is proportional to either 10%, 50%, or 90% (Figure 4A–4C).

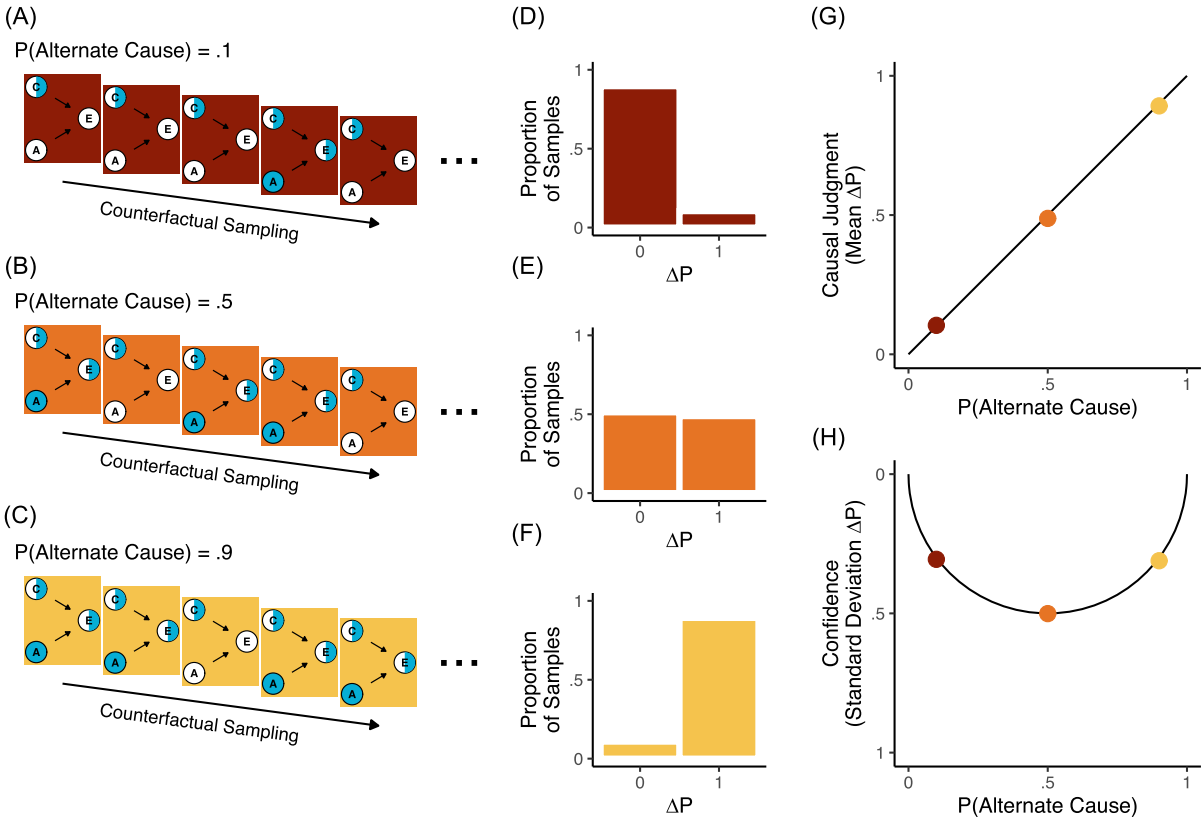
Next, counterfactual sampling models predict that people evaluate, separately in each possibility, whether Joe drawing the green ball (or not) made a difference to him winning the dollar (or not). Specifically, within each possibility, people manipulate whether Joe drew the green ball or not (*C*), and they check whether Joe draws the dollar (*E*; represented by half-filled nodes in Figure 4A–4C). In this causal structure, manipulating whether Joe draws a green ball only effects whether he won the dollar when he also drew a blue ball (*A*) because without drawing a blue ball, there is no way for drawing the green ball to help him win the dollar. So, graphically, ΔP predicts that drawing a green ball makes a difference to winning the dollar precisely when the node for the effect *E* is half-filled (Figure 4D–4F).

Finally, counterfactual sampling models predict that people compute summary statistics over the distribution of specific causal effects to form different judgments. Specifically, their reported causal judgment is the mean of the distribution (depicted as points in Figure 4G). We can see that as the probability that Joe draws a blue ball from the right box increases, so do causal judgments of the green ball because according to the ΔP model, Joe drawing the green ball can only be said to make a difference to winning the dollar

[3] It can be argued that the standard error provides a more normative measure of precision because it takes into account how many counterfactual samples were considered when making a judgment. Indeed, previous work suggests that people only consider a small number of possibilities in similar tasks (Phillips et al., 2019; Vul et al., 2014). However, it is experimentally difficult to determine how many samples one considers when making a judgment, precluding us from using this information in our model predictions. Additionally, we have no reason to expect that our manipulations of causal structure and statistical normality influence the number of samples participants consider, meaning that this parameter is likely constant across our experiment, and so the standard error and the standard deviation yield the same predictions in this context. So, while we do not explore this measure in the current article, future work may dissociate these competing hypotheses.

Figure 4

The Counterfactual Sampling Model of Normality Effects on Causal Judgment and Confidence



Note. (A)–(C) Depiction of the counterfactual sampling process when the probability of the alternate cause *A* is either (A) .1, (B) .5, or (C) .9. Each box depicts an imagined possibility where events with filled nodes occur and those with unfilled nodes do not occur. Half-filled nodes depict that within each possibility, the *C* is intervened upon to see whether it makes a difference to the *E*. (D)–(F) Histograms of the difference the candidate cause made to the effect predicted by ΔP when the probability of the alternate cause is either (D) .1, (E) .5, or (F) .9. (G) Predicted causal judgment using mean ΔP as a function of the probability of the alternate cause. (H) Predicted confidence using the standard deviation of ΔP as a function of the probability of the alternate cause. The *y*-axis is reversed so that smaller standard deviations indicate higher confidence. *E* = effect; *C* = candidate cause; *A* = alternate cause. See the online article for the color version of this figure.

when he draws the blue ball. This increase in causal judgments of the candidate cause when the normality of the alternate cause increases is known as causal superseding and has been observed in empirical data (Kominsky et al., 2015; Morris et al., 2019).

According to our precision model of confidence, however, people can report upon more than just the mean of this distribution. Notably, they can report their confidence using a measure of precision. The standard deviations of the corresponding distributions are depicted in Figure 4H. Here, we see a different pattern than for the means: People are expected to be confident when the probability of drawing the blue ball is low (Figure 4A) or high (Figure 4C), but they are expected to be uncertain when this probability is close to 50% (Figure 4B). This is because drawing the green ball has a much more variable effect on winning the dollar when the probability of drawing the blue ball is 50%. Without any information about whether Joe would have drawn a blue ball, it is difficult to predict whether or not drawing the green ball would make a difference to winning the dollar. When the probability of drawing a blue ball is low, drawing the green ball rarely makes a difference to winning the dollar, so one can be confident that the green ball did not cause Joe to win the dollar. Likewise, when the probability of drawing a blue ball is high, drawing the green ball almost always makes a difference to winning the dollar, so one can be confident that the green ball did cause Joe to win the dollar. Overall, then, we have the U-shaped relationship between predicted causal judgment and predicted confidence depicted in Figure 3: People should be least confident when their causal judgment is close to the midpoint of the scale.

The Current Article

Our model of confidence in causal judgment is a conjunction of counterfactual sampling models of causal judgment and precision models of confidence: Given a distribution of differences the candidate cause could have made to the effect, causal judgments are reports of the average difference made by the candidate cause, and

confidence ratings are reports of the variation in these differences. To test this model, we replicated and extended a recent study measuring quantitative shifts in causal judgments with respect to the probabilities of the candidate and alternate causes, $P(C)$ and $P(A)$ (Morris et al., 2019). Previous work has shown that causal judgments of C tend to decrease with $P(C)$ but increase with $P(A)$ in conjunctive causal structures and that they increase with $P(C)$ but decrease with $P(A)$ in disjunctive causal structures (Gill et al., 2022; Icard et al., 2017; Kominsky et al., 2015; Kominsky & Phillips, 2020; Morris et al., 2019). Because each of the above measures of uncertainty predicts that people's confidence in their causal judgments is a function of the causal judgments themselves, they predict that confidence should also vary with $P(C)$ and $P(A)$: Specifically, confidence should be high when causal judgments are predicted to be very high or low, and confidence should be low when causal judgments are predicted to be intermediary (Figure 3). Accordingly, we also measure participants' confidence in their causal judgments.

Our study thus provides two tests of different assumptions made by counterfactual sampling models of causal judgment, visualized as arrows in Figure 2. First, as in Morris et al.'s (2019) study, we use causal judgments to evaluate whether each model describes the kind of difference making relevant to causal judgment. So, if a particular model fails to predict causal judgments, we would have reason to believe that another notion of difference making is required to account for this pattern in judgments (Figure 2, dashed arrow a).

Second, given that a model predicts causal judgments, we use confidence ratings to test the other assumptions made by the counterfactual sampling framework. If a model predicts causal judgments but not confidence ratings, then although we have evidence that this notion of difference making generally captures people's causal judgments, an auxiliary assumption must be amended to predict confidence. One possibility is that the precision account is wrong and that confidence ratings are actually nonnormative: Even if people sample many possibilities to ensure that the estimated counterfactual effect is robust to changes in background circumstances, they discard this information when rating confidence (dashed arrow b in Figure 2). This interpretation, however, would require a strong justification for why people would ignore information about the robustness of a causal effect after just employing an intensive and normative procedure to generate this information when making a causal judgment. Instead, we advocate for an alternative interpretation whereby confidence ratings can be used to determine whether people use Monte Carlo sampling to simulate alternative possibilities (dashed arrow c in Figure 2). So, if a model predicts causal judgments but not confidence ratings, we have reason to doubt that people come to this judgment using the particular pattern of sampling typically assumed by counterfactual sampling models. In this case, revising this model would require proposing an alternative sampling mechanism with similar predictions of causal judgment but different predictions of confidence (dashed arrow c in Figure 2).

Finally, if any models are able to jointly predict causal judgments and confidence ratings across our experimental conditions, this would provide strong evidence both for the measure of difference making and for the broader mechanism of counterfactual sampling in causal judgment because these measures were developed solely to predict causal judgments and not confidence. In sum, we compared model predictions of causal judgments to empirical judgments to test the measures of difference making assumed by different counterfactual models, and we compared model predictions of the precision of counterfactual effects with empirical confidence ratings to test the specific commitment to sampling made by these models.

Method

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study, and we follow Journal Article Reporting Standards (Kazak, 2018). All data, analysis code, and research materials are available at https://osf.io/sm6qg/. Data were analyzed using R, Version 4.1.2 (R Core Team, 2021), and the cmdstanr interface to the probabilistic programming language Stan (Carpenter et al., 2017; Gabry & Češnovar, 2021). This study's design and its analysis were not preregistered.

Participants

Based on the sample size from Morris et al. (2019), along with the expectation that normality effects on confidence would be smaller than effects on causal judgment, we recruited 3,020 participants from Prolific (https://prolific.co). All participants were from the United States, spoke English as their native language, and provided informed consent in accordance with Duke University's institutional review board. The participants completed the task in an average of 7.5 min and were compensated with $0.75. One hundred eighteen (3.9%) participants were excluded from our analyses because they reported not paying attention to the task in response to an explicit attention check after completion of the task (see Supplemental Material). Data were analyzed from the remaining 2,902 participants ($M_{age} = 36.93$, $SD_{age} = 13.23$). One thousand four hundred twenty-one participants identified their gender as female, 1,444 as male, 35 as other, and two participants chose not to report their gender.

Materials

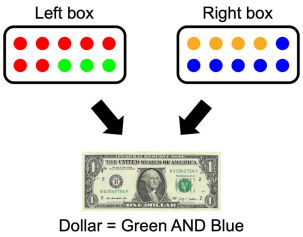
Stimuli were six vignettes similar to the vignette used in Morris et al.'s (2019; see Supplemental Material) study. Each vignette included a deterministic causal system involving two candidate causes (which could occur independently with defined probabilities) and an outcome that would occur if and only if both candidate causes occurred (conjunctive structure) or if either candidate cause occurred (disjunctive structure). In all vignettes, the two candidate causes always occurred, and so the outcome also always occurred. The outcome was positive (e.g., winning a dollar) in half of the vignettes and negative (e.g., having to pay for drinks) in the other half. Alongside each vignette, the participants were shown an image that briefly summarized the vignette and also defined the probability of each candidate cause. All stimuli, materials, and code are accessible via the Open Science Framework (https://osf.io/sm6qg/). An example stimulus is depicted in Figure 5.

Procedure

In a $10 \times 10 \times 2 \times 6$ sparsely sampled within-participants design (probability of candidate cause: {.1, .2, … 1}; probability of alternate cause: {.1, .2, … 1}; causal structure: {Conjunctive, Disjunctive}; vignette), participants read one version of each of the

Figure 5

Example Stimulus Presented to the Participants



A person, Joe, played a casino game where he reached into two boxes and blindly drew a ball from each box.	
Conjunctive: In this game, he wins a dollar if and only if he gets a green ball from the left box and a blue ball from the right box. If he doesn't get a green ball from the left box or he doesn't get a blue ball from the right box, he doesn't win a dollar.	**Disjunctive:** In this game, he wins a dollar if he gets a green ball from the left box or a blue ball from the right box (or both). If he doesn't get a green ball from the left box and he doesn't get a blue ball from the right box, he doesn't win a dollar.
Joe closed his eyes, reached a hand into each box, and chose a green ball from the left box and a blue ball from the right box. So Joe won the dollar.	
To what degree did Joe win the dollar because he drew a green ball from the left box? **How confident are you in your response to the previous question?**	

Note. The participants were asked to read one version of the vignette (conjunctive or disjunctive). The image was presented along with the vignette and contained information about the probabilities of the candidate and alternate causes. In this example, the probability of the candidate cause (drawing a green ball) is .3, and the probability of the alternate cause (drawing a blue ball) is .6. See the online article for the color version of this figure.

six vignettes (participants each read six vignettes total). The probability of each candidate cause and the causal structure was randomly assigned for each vignette, and the order of vignettes was randomized. For each vignette, participants read the vignette and inspected a corresponding image, which added information about the probability of each event occurring. The participants then responded to the questions "To what degree did [the effect occur] because [the candidate cause occurred]?" and "How confident are you in your response to the previous question?" on continuous slider scales ranging from *not at all* (coded as 0) to *totally* (coded as 1).

Analysis

As a descriptive model of the effects of the normality of the candidate and alternate causes on both causal judgments and confidence ratings, we fit a bivariate Gaussian process (GP) model using the probabilistic programming language Stan (Carpenter et al., 2017; Stan Development Team, 2020, 2021). The GP has two main advantages as a statistical model. First, it is known that normality effects on causal judgments and confidence are in fact nonlinear (Morris et al., 2019), and the GP can capture such nonlinear relationships. Second, it assumes that effects are smooth, which helps to penalize overfitting and reduce statistical errors (Rasmussen & Williams, 2005). We estimated the mean and variability of causal judgments and confidence ratings with separate latent GPs for conjunctive and disjunctive causal structures, including a vignette-specific GP to account for vignette-level effects (see Supplemental Material for mathematical details). Specifically, the GPs jointly modeled the mean and precision parameters of an ordered Beta likelihood using a logit and log link function, respectively (Kubinec, 2020), which accounts for the fact that both causal judgments and confidence ratings were bounded between 0 and 1 with many responses at precisely these bounds. To test for changes with respect to the probability of each cause, we also jointly estimated the gradients of each GP with respect to the probabilities of the candidate and alternate causes, and we report the largest gradients as β values (Riihimäki & Vehtari, 2010; Solak et al., 2003). For each parameter, we report the posterior median and 95% highest density intervals (HDIs). We used a Bayesian analog of the p value computed from the probability of direction with a threshold of .05 to test for effect existence (Makowski et al., 2019), and we considered any parameter with a Bayes Factor (BF) greater than 10 as statistically significant. For supplementary results, vignette-level effects, and model diagnostics, see Supplemental Material.

To compare the predictions of the different counterfactual sampling models, we fitted them each to participants' causal judgments and confidence ratings as generative models in Stan. For simplicity, past work has assumed that the subjective counterfactual sampling probability of the candidate and alternate causes is equal to their objective probabilities (Morris et al., 2019; Quillien, 2020). Here, we relaxed this assumption by instead assuming that the subjective counterfactual sampling probabilities were positive monotonic functions of the objective probabilities, which could vary by causal structure (see Supplemental Material for fitted counterfactual sampling probabilities). Because we did not want to assume that the model predictions were on the same scale as participants' judgments, we assumed a linear relationship between the mean counterfactual difference and mean causal judgments. Similarly, we also assumed a linear relationship between the precision of counterfactual differences and mean confidence. For each model, we report the estimated mean causal judgment and confidence rating per condition. For formal model comparisons of generative performance, we used the approximate leave-one-out cross-validated expected log pointwise predictive density (ELPD_LOO) separately for causal judgments and confidence, which evaluates the ability of each model to predict causal judgments and confidence ratings on held-out data (Vehtari et al., 2017).

Results

Causal Judgment

We first sought to replicate previous results showing that causal judgments vary as a function of the probability of the candidate cause (i.e., the cause that we ask participants to judge) and the alternate cause (i.e., the cause that participants do not judge; Icard et al., 2017; Kominsky et al., 2015; Morris et al., 2019). Figure 6A and 6B depicts mean causal judgments and predictions from each model, respectively. In conjunctive structures, causal judgments of the candidate cause tended to decrease with the probability of the candidate cause ($\beta = -.55$, 95% HDI [−.93, −.21], $P < .001$, BF = 1,294) and increase with the probability of the alternate cause ($\beta = .19$, 95% HDI [.06, .33], $P < .001$, BF = 1,369). In disjunctive structures, causal judgments tended to increase with the probability of the candidate cause ($\beta = .10$, 95% HDI [.004, .20], $P = .04$, BF = 20) and decrease with the probability of the alternate cause ($\beta = -.08$, 95% HDI [−.17, −.01], $P = .01$, BF = 77).

We then asked whether these patterns in causal judgments were predicted by counterfactual models. To answer this question, we computed the approximate ELPD_LOO (Vehtari et al., 2017) for each model, and we computed differences of model performance relative to the best performing model (Figure 7, left panel). Here, we report the performance of the best-performing measure of precision for each measure of difference making, though results were similar across different measures of precision. As found in previous works (Morris et al., 2019; Quillien, 2020), counterfactual sampling models were largely successful in predicting normality effects on causal judgments. In particular, the Counterfactual Effect Size model had the best predictions of causal judgments (ELPD_LOO = −11751.03, *SE* = 103.26), although the Necessity–Sufficiency model (ELPD_LOO = −11757.19, *SE* = 104.19, ΔELPD_LOO = −6.16, *SE* = 9.12) and the Crediting Causality model (ELPD_LOO = −11758.35, *SE* = 103.13, ΔELPD_LOO = −7.32, *SE* = 4.95) did not make significantly worse predictions. The Power PC model (ELPD_LOO = −11803.07, *SE* = 104.22, ΔELPD_LOO = −52.04, *SE* = 12.47) and the Δ*P* model (ELPD_LOO = −11781.13, *SE* = 104.25, ΔELPD_LOO = −30.10, *SE* = 10.04) performed significantly worse than the Counterfactual Effect Size model in predicting causal judgments.

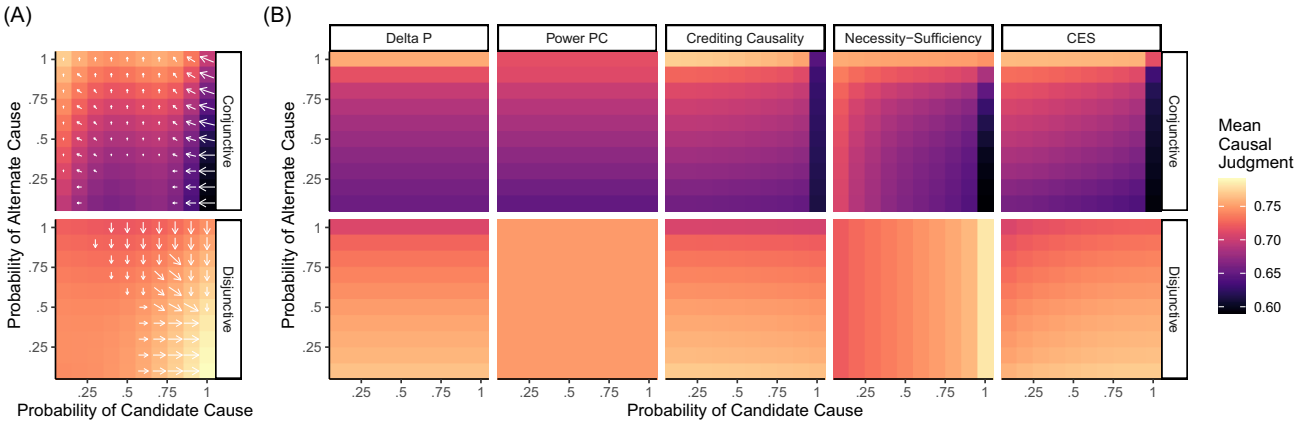
Confidence

Next, we asked whether participants' confidence in their causal judgments also varied with respect to the probability of the candidate and alternate causes. Figure 8A and 8B depicts mean confidence and predictions from each model, respectively. In conjunctive structures, participants tended to be less confident in their causal judgments with increases in the probability of the candidate cause ($\beta = -.09$, 95% HDI [−.18, −.02], $P = .02$, BF = 30) and more confident with increases in the probability of the alternate cause ($\beta = .06$, 95% HDI [.01, .12], $P = .01$, BF = 30). There was also a small region in which participants were more confident with increases in the probability of the candidate cause ($\beta = .06$, 95% HDI [.01, .12], $P = .04$, BF = 14), indicating that normality effects on confidence are likely nonmonotonic. In contrast, in disjunctive structures, participants tended to be more confident as the probability of the candidate cause increased ($\beta = .11$, 95% HDI [.01, .22], $P = .03$, BF = 28), though there was no effect of the probability of the alternate cause ($\beta = .03$, 95% HDI [−.04, .09], $P = .29$, BF = 3). White arrows in Figure 8 depict regions where these effects were significant. However, we note that confidence was very high overall ($M = .84$, $SD = .22$) and that the observed effects on confidence were small compared with the corresponding effects on causal judgment.

Finally, we tested whether precision models of confidence, in conjunction with counterfactual sampling models of causal judgment, predicted participants' confidence in their causal judgments. Figure 7 (right panel) depicts the performance of each model in predicting confidence ratings. Here, the Necessity–Sufficiency model had the best predictions of confidence ratings (ELPD_LOO = −9703.16, *SE* = 90.70). The Counterfactual Effect Size model

Figure 6

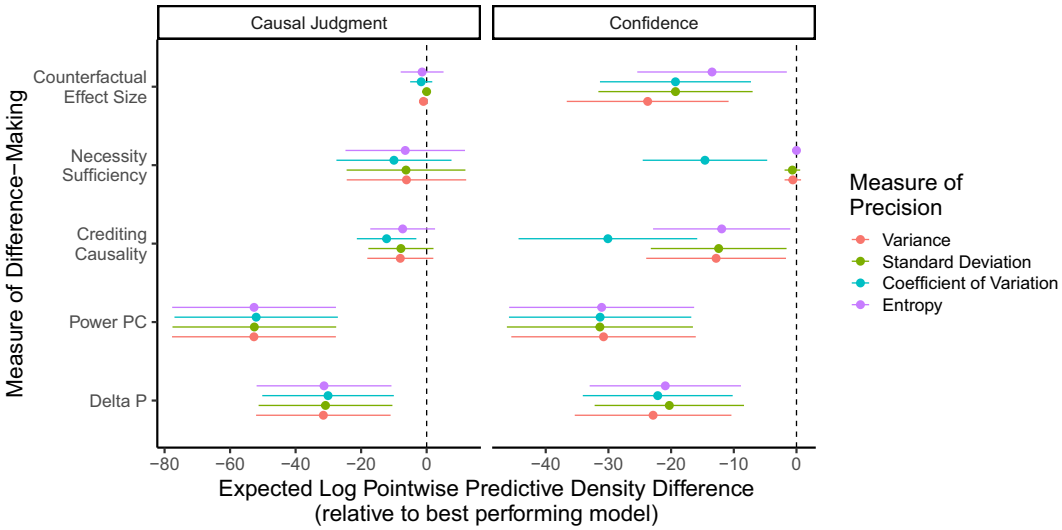
Inferred Mean Causal Judgment (A) Compared With Model Predictions Using the Standard Deviation Measure of Precision (B)



Note. Arrows indicate significant effects on mean causal judgment with respect to the probability of the candidate or alternate causes. The length of the arrow is proportional to the size of the effect, and the arrows point in the direction of increasing mean causal judgment. PC = probabilistic contrast; CES = Counterfactual Effect Size. See the online article for the color version of this figure.

Figure 7

Relative Model Performance for Causal Judgments and Confidence Ratings



Note. Values close to 0 indicate that the model performs as well as the best model in predicting causal judgments or confidence, and lower values indicate worse performance. While most models perform similarly well at predicting causal judgments, only the Necessity–Sufficiency model predicts causal judgments and confidence. Points indicate means, and error bars indicate twice the standard error. PC = probalistic contrast. See the online article for the color version of this figure.

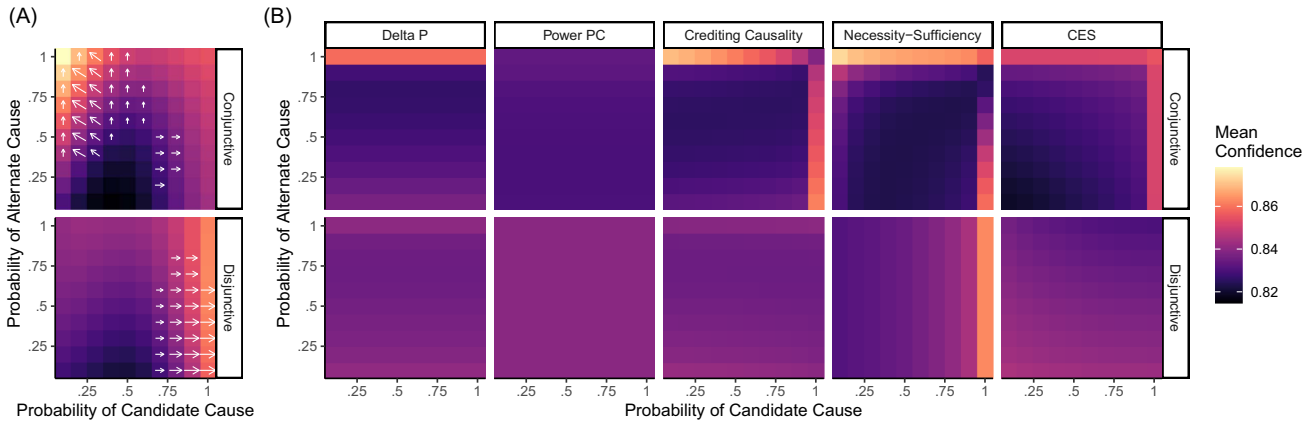
(ELPD_LOO = −9716.63, *SE* = 90.74, ΔELPD_LOO = −13.47, *SE* = 5.96), the Crediting Causality model (ELPD_LOO = −9715.09, *SE* = 90.68, ΔELPD_LOO = −11.93, *SE* = 5.47), the Power PC model (ELPD_LOO = −9733.94, *SE* = 90.28, ΔELPD_LOO = −30.78, *SE* = 7.36), and the Δ*P* model (ELPD_LOO = −9723.45, *SE* = 90.58, ΔELPD_LOO = −20.29, *SE* = 5.96) all made significantly worse predictions of confidence. In sum, the Necessity–Sufficiency model outperformed all other models in predicting confidence in causal judgments.

Relationships Between Causal Judgments and Confidence

As a secondary hypothesis, we also sought to determine whether there were changes in the estimated variability of causal judgments (Figure 9A) and confidence (Figure 9B). In conjunctive structures, variance in causal judgments of the candidate cause tended to increase with the probability of the candidate cause (β = .15, 95% HDI [.08, .23], *P* < .001, BF = 2003) and decrease with the probability of the alternate cause (β = −.04, 95% HDI [−.08, −.01],

Figure 8

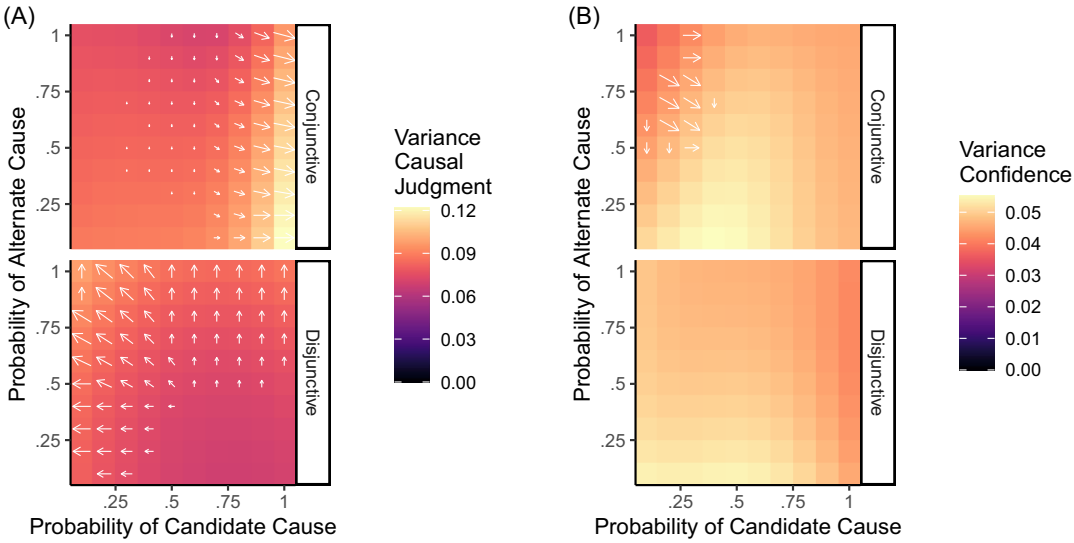
Mean Confidence in Causal Judgment (A) Compared With Model Predictions Using the Standard Deviation Measure of Precision (B)



Note. Arrows indicate significant effects on mean confidence with respect to the probability of the candidate or alternate causes. The length of the arrow is proportional to the size of the effect, and the arrows point to the direction of increasing mean confidence. PC = probabilistic contrast; CES = Counterfactual Effect Size. See the online article for the color version of this figure.

Figure 9

Normality Effects on the Variance of Causal Judgments and Confidence Ratings



Note. (A) Estimated variance of causal judgments. (B) Estimated variance of confidence ratings. For visibility, color scales differ between the two plots. Arrows indicate significant effects on the variance with respect to the probability of the candidate or alternate causes. The length of the arrow is proportional to the size of the effect, and the arrows point in the direction of increasing variance. See the online article for the color version of this figure.

P = .006, BF = 99). Variance in confidence ratings also tended to increase with the probability of the candidate cause (β = .03, 95% HDI [.001, .06], P = .04, BF = 39) and decrease with the probability of the alternate cause (β = −.02, 95% HDI [−.03, −.001], P = .02, BF = 14). In disjunctive structures, variance of causal judgments tended to decrease with the probability of the candidate cause (β = −.04, 95% HDI [−.09, −.001], P = .04, BF = 20) and increase with the probability of the alternate cause (β = .03, 95% HDI [.006, .07], P = .007, BF = 83), though we found little evidence for effects of the probability of either cause on the variance of confidence ratings (β = −.03, 95% HDI [−.07, .01], P = .12, BF = 9).

Overall, the mean and variance were strongly negatively correlated for both causal judgments (r = −.87, 95% HDI [−.93, −.81], P < .001, BF > 10,000) and confidence (r = −.96, 95% HDI [−.99, −.91], P < .001, BF > 10,000). There was weak evidence for a correlation between mean causal judgment and confidence (r = .31, 95% HDI [.002, .60], P = .04, BF = 4), but there was no evidence for a correlation between the variance of causal judgments and mean confidence (r = −.03, 95% HDI [−.38, .26], P = .81, BF = .44). So, even though the variance of causal judgments and mean confidence showed qualitatively similar trends, it was not the case that participants agreed with each other more when they were more confident about their causal judgments.

Discussion

In this article, we proposed an extension of counterfactual sampling models of causal judgment to include confidence in those judgments. Our extension, following recent work in metacognition, is simple: While people's causal judgments are explained by the average difference the candidate cause is thought to make to the effect, their confidence is the precision of the distribution around this estimate, using, for instance, the inverse standard deviation (Ma & Jazayeri, 2014; Meyniel & Dehaene, 2017; Meyniel et al., 2015; Navajas et al., 2017; Pouget et al., 2016; Yeung & Summerfield, 2012). Our model made the novel prediction that people should be more or less confident in their causal judgments depending on the probability of each of the causes. To test different variations of the model, we replicated and extended an experiment by Morris et al. (2019). We found that participants' confidence decreased with the probability of the candidate cause and increased with the probability of the alternate cause in conjunctive causal structures, whereas their confidence increased with the probability of the candidate cause in disjunctive causal structures. Critically, these patterns were best predicted by a single model: the Necessity–Sufficiency model. Because the Necessity–Sufficiency model was developed solely to explain causal judgments (with no regard for confidence), our results provide strong support for this model.

In contrast, all of the other counterfactual sampling models provided significantly worse predictions of either causal judgments or confidence ratings. Although the Counterfactual Effect Size model predicted causal judgments well in both structures, it failed to predict confidence in disjunctive causal structures: It predicted that confidence would decrease with the probability of the candidate cause when it in fact increased. Similarly, the Crediting Causality model predicted causal judgments well in conjunctive structures, but not in disjunctive structures, where it predicted that judgments would decrease with the probability of the candidate cause when they actually increased. In conjunctive structures, it also predicted that confidence would increase with increases in the probability of the candidate cause and decreases in the probability of the alternative cause, which is the opposite of what was actually found. The Power PC model was able to predict causal judgments and confidence in conjunctive structures, but it predicted no changes in either rating in

disjunctive structures. The ΔP model significantly predicted causal judgments in both causal structures, but it also predicted no effect of the probability of the candidate cause on confidence in disjunctive structures. Thus, each model predicted some features of causal judgments and confidence in those judgments while failing to predict others.

It may be tempting to conclude, then, that our results provide decisive evidence for the Necessity–Sufficiency model and against the other counterfactual sampling models. To the contrary, we found that the Necessity–Sufficiency model had its own explanatory gap: Notably, it did not predict our replication of Morris et al.'s (2019) findings that people's causal judgments decreased with the probability of the alternate cause in disjunctive structures. Moreover, we also replicated past findings that all models were successful at predicting causal judgments in conjunctive structures, and all but the Crediting Causality and the Power PC models were successful in disjunctive structures (Morris et al., 2019).

How should we revise our models of causal judgment in light of these varied results? To answer this question, we return to our theoretical framework introduced in Figure 2. First, we can use causal judgments as a test of which particular kind of difference making is relevant to causal judgment (Figure 2, dashed arrow a). While none of the models provided perfect predictions, most of the model predictions were quantitatively similar to the observed patterns in causal judgments, making it difficult to provide clear evidence against them. Next, because models that give similar predictions of causal judgments give qualitatively different predictions of confidence (and in fact most models failed to predict confidence), we can use confidence ratings as a secondary test of these models. Of course, one interpretation is that confidence ratings, like causal judgments, directly reflect which kind of difference making is relevant to causal judgment, so only the Necessity–Sufficiency model is acceptable. But given that other measures do predict causal judgments, we think it is premature to dismiss them entirely on the basis of the confidence ratings alone. Another possibility is that these models failed to predict confidence because our precision account is simply false (Figure 2, dashed arrow b), and confidence is estimated using some other procedure such as a nonnormative heuristic (Adler & Ma, 2018) or a more complex second-order inference (Fleming & Daw, 2017). As we argue in the Introduction section, however, this interpretation would require a justification of why people would use an intensive sampling procedure when making a causal judgment only to ignore this information when making metacognitive assessments. Moreover, we found that—as predicted by our normative account—confidence ratings were sensitive to normality and quadratically related to causal judgments, making it unlikely that our participants simply had poor metacognitive sensitivity.

Instead, we advocate for a third interpretation whereby confidence ratings reflect whether counterfactual effects are computed using a standard Monte Carlo sampling scheme (Figure 2, dashed arrow c). So, the models that failed to predict confidence could be revised to make use of a different pattern sampling. For instance, existing counterfactual sampling models assume that people simulate a sufficiently large number of counterfactual possibilities to estimate counterfactual effects and that this sample size is relatively constant across contexts. This sampling scheme could be modified by assuming that people only simulate a small number of possibilities due to time pressure or general capacity limits (Phillips et al., 2019; Vul et al., 2014), that the number of possibilities simulated depends on contextual factors, or that people are more inclined to consider possibilities similar to what actually happened (Lucas & Kemp, 2015; Quillien & Lucas, 2023). In any case, it is clear that future work will have to explore these possibilities to converge upon a unified account of both causal judgments and confidence.

One limitation with the present study is that in focusing on counterfactual sampling models, we were unable to provide direct evidence that counterfactual sampling models outperform other kinds of models of confidence in causal judgments. For example, process theories propose that causal judgments reflect assessments of the transmission of physical quantities like force from the cause to the effect (Wolff, 2007), and social cognitive theories propose that causal judgments reflect prior ascriptions of blame or responsibility (Alicke et al., 2011; Sytsma, 2021). There remains debate between proponents of these different theories (Henne & O'Neill, 2022; Kominsky & Phillips, 2020; Krasich et al., 2024). However, we still interpret our results in favor of counterfactual theories because, to our knowledge, they are the only theories that are currently capable of making quantitative predictions of normality effects on causal judgments, so they are also the only theories capable of predicting such effects on confidence. Although we found that people make causal judgments and confidence ratings in a way that is consistent with counterfactual sampling models, and although participants' inferred counterfactual sampling probabilities strongly resemble representations of probability in other domains (see Supplemental Material; Tversky & Kahneman, 1992; Zhang & Maloney, 2012), another limitation is that we did not directly measure counterfactual judgments or confidence in counterfactual judgments. So, following other research demonstrating counterfactual representations during causal judgment (Gerstenberg et al., 2017; Henne & O'Neill, 2022; Krasich et al., 2024), future work could look for similar indices of counterfactual representations when rating confidence.

More work is needed to investigate metacognitive assessments of causal judgments in more ecologically valid domains. In our task, participants had full information about the relevant variables, the causal structure, and the actual events that took place, so they reported very high confidence overall. This design had the advantage that any uncertainty reported by participants necessarily comes internally from within the participants themselves, allowing us to isolate that uncertainty as coming from the hypothesized sampling process. But people most often make causal judgments in the presence of these types of uncertainty in addition to mere probabilistic uncertainty. It is also widely known that metacognition of perceptual and value-based decisions affects learning, exploration, and changes of mind (Folke et al., 2016; Kepecs et al., 2008; Shea et al., 2014). Future work, then, should explore the ways in which metacognition about causal judgments impacts subsequent cognition. For instance, despite differences between the problems of singular and general causal judgment, it is unclear whether people might use similar strategies in rating confidence for the two kinds of judgment (Liljeholm, 2015). It is also unclear whether varying degrees of confidence in causal judgments might impact decision making in other tasks involving causal representations such as one-shot learning, reversal learning, and reward-based learning (Chambon et al., 2018; Lehmann et al., 2019; Rouault et al., 2022; Weiss et al., 2021). Finally, future research may investigate the relationship between confidence and causal judgments in other causal structures. We replicated previous findings that mean causal judgment and confidence are quadratically related (Liljeholm, 2015; O'Neill,

Henne, et al., 2022), and our precision model of confidence provides one explanation for this relation: Both judgments reflect different summary statistics of the same underlying distribution of counterfactual differences. When the effect is binary, such differences are Bernoulli-distributed, which necessarily implies a U-shaped relation between causal judgments (see Figure 3). So, future work may benefit from extending counterfactual models to cases where this relation between confidence and causal judgments need not hold (e.g., where the effect is a continuous variable), allowing for an empirical dissociation of whether confidence in causal judgments are sensitive to uncertainty in the estimate of the mean counterfactual difference, variability in the distribution of counterfactual differences, or both.

Constraints on Generality

In this article, we recruited English-speaking participants from the United States and presented them with text-based stimuli. However, we would expect normality effects on causal judgments to generalize to participants from various social groups, as well as to experiments with other kinds of stimuli (see Gerstenberg & Icard, 2020; Henne & O'Neill, 2022; Henne, O'Neill, et al., 2021, for examples with video-based stimuli). We manipulated statistical norms by varying the probability of different events, but similar findings have also been found with prescriptive or social norms that are presumably more sensitive to cultural factors (e.g., Güver & Kneer, 2023; Icard et al., 2017). Finally, we note that the predicted quadratic relationship between mean causal judgment and mean confidence is specific to causal structures with binary variables, and so this relationship should not be expected to generalize to cases with continuous variables.

Conclusion

In sum, we proposed an extension of counterfactual sampling models of human causal judgment to additionally predict confidence in those judgments, allowing us to use confidence ratings as a test of the sampling mechanism underlying recent counterfactual models of causal judgment. When compared with judgments made by participants, one version of this model (using the Necessity–Sufficiency measure of causal strength) was able to simultaneously predict causal judgments and confidence in those judgments (Icard et al., 2017). Our results, in addition to furthering our understanding of causal judgment, are an important step in determining the mechanisms behind metacognitive assessments of complex decisions.

References

Adler, W. T., & Ma, W. J. (2018). Comparing bayesian and non-bayesian accounts of human confidence reports. *PLOS Computational Biology*, *14*(11), Article e1006572. https://doi.org/10.1371/journal.pcbi.1006572

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *The Journal of Philosophy*, *108*(12), 670–696. https://www.jstor.org/stable/23142912

Beebee, H., Hitchcock, C., & Menzies, P. (2009). *The Oxford handbook of causation*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199279739.001.0001

Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, *35*(8), 3478–3484. https://doi.org/10.1523/JNEUROSCI.0797-14.2015

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1880–1910. https://doi.org/10.1037/xlm0000548

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Chambon, V., Thero, H., Findling, C., & Koechlin, E. (2018). *Believing in one's power: A counterfactual heuristic for goal-directed control*. bioRxiv. https://doi.org/10.1101/498675

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405. https://doi.org/10.1037/0033-295X.104.2.367

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*(4), 545–567. https://doi.org/10.1037/0022-3514.58.4.545

Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to white (2005) and to luhmann and ahn (2005). *Psychological Review*, *112*(3), 694–706. https://doi.org/10.1037/0033-295X.112.3.694

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*, 93–115. https://doi.org/10.1613/jair.1391

Danks, D. (2017). Singular causation. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 201–215). Oxford University Press.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. https://doi.org/10.1038/nn.3279

Demirtas, H. (2022). Causation comes in degrees. *Synthese*, *200*(1), Article 64. https://doi.org/10.1007/s11229-022-03507-2

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science*, *29*(5), 761–778. https://doi.org/10.1177/0956797617744771

Dotan, D., Meyniel, F., & Dehaene, S. (2018). On-line confidence monitoring during decision making. *Cognition*, *171*, 112–121. https://doi.org/10.1016/j.cognition.2017.11.001

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045

Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, *1*(1), Article 0002. https://doi.org/10.1038/s41562-016-0002

Gabry, J., & Češnovar, R. (2021). *Cmdstanr: R interface to 'cmdstan'*. https://mc-stan.org/cmdstanr

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936–975. https://doi.org/10.1037/rev0000281

Gerstenberg, T., & Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599–607. https://doi.org/10.1037/xge0000670

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744. https://doi.org/10.1177/0956797617713053

Gill, M., Kominsky, J., Icard, T., & Knobe, J. (2022). An interaction effect of norm violations on causal judgment. *Cognition*, *228*, Article 105183. https://doi.org/10.1016/j.cognition.2022.105183

Godfrey-Smith, P. (2009). Causal pluralism. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford handbook of causation* (pp. 326–338). Oxford University Press.

Goupil, L., Romand-Monnier, M., & Kouider, S. (2016). Infants ask for help when they know they don't know. *Proceedings of the National*

Academy of Sciences, *113*(13), 3492–3496. https://doi.org/10.1073/pnas.1515129113

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384. https://doi.org/10.1016/j.cogpsych.2005.05.004

Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., & Usher, M. (2020). Causal responsibility and robust causation. *Frontiers in Psychology*, *11*, Article 1069. https://doi.org/10.3389/fpsyg.2020.01069

Güver, L., & Kneer, M. (2023). Causation and the silly norm effect. In S. Magen & K. Prochownik (Eds.), *Advances in experimental philosophy of law* (pp. 133–168). Bloomsbury Publishing.

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of their choices as interventions. *Journal of Experimental Psychology: General*, *138*(1), 22–38. https://doi.org/10.1037/a0014585

Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science*, *66*(2), 413–457. https://doi.org/10.1093/bjps/axt050

Hangya, B., Sanders, J. I., & Kepecs, A. (2016). A mathematical framework for statistical decision confidence. *Neural Computation*, *28*(9), 1840–1858. https://doi.org/10.1162/NECO_a_00864

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

Henne, P. (2023). Experimental metaphysics: Causation. In A. Bauer & S. Kornmesser (Eds.), *The compact compendium of experimental philosophy* (pp. 133–161). De Gruyter.

Henne, P., Kulesza, A., Perez, K., & Houcek, A. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition*, *212*, Article 104708. https://doi.org/10.1016/j.cognition.2021.104708

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. https://doi.org/10.1016/j.cognition.2019.05.006

Henne, P., & O'Neill, K. (2022). Double prevention, causal judgments, and counterfactuals. *Cognitive Science*, *46*(5), Article e13127. https://doi.org/10.1111/cogs.13127

Henne, P., O'Neill, K., Bello, P., Khemlani, S., & De Brigard, F. (2021). Norms affect prospective causal judgments. *Cognitive Science*, *45*(1), Article e12931. https://doi.org/10.1111/cogs.12931

Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing ourselves together: The cultural origins of metacognition. *Trends in Cognitive Sciences*, *24*(5), 349–362. https://doi.org/10.1016/j.tics.2020.02.007

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*(5), 942–951. https://doi.org/10.1086/667899

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*(11), 587–612. https://www.jstor.org/stable/20620209

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with bayesian causal models. *Journal of Experimental Psychology: General*, *139*(4), 702–727. https://doi.org/10.1037/a0020488

Hume, D. (1748). *Philosophical essays concerning human understanding*. A. Millar.

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

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1–17. https://doi.org/10.1037/h0093874

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153. https://doi.org/10.1037/0033-295X.93.2.136

Kaiserman, A. (2016). Causal contribution. *Proceedings of the Aristotelian Society*, *116*(3), 387–394. https://doi.org/10.1093/arisoc/aow013

Kaiserman, A. (2018). 'More of a cause': Recent work on degrees of causation and responsibility. *Philosophy Compass*, *13*(7), Article e12498. https://doi.org/10.1111/phc3.12498

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322–1337. https://doi.org/10.1098/rstb.2012.0037

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231. https://doi.org/10.1038/nature07200

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764. https://doi.org/10.1126/science.1169405

Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, Article 104721. https://doi.org/10.1016/j.cognition.2021.104721

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 2. The cognitive science of morality: Intuition and diversity* (pp. 441–447). Boston Review.

Kominsky, J. F., & Phillips, J. (2020). "Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection": Erratum. *Cognitive Science*, *43*(11), Article e12792. https://doi.org/10.1111/cogs.12821

Kominsky, J. F., 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

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), Article e13426. https://doi.org/10.1111/cogs.13426

Kubinec, R. (2020). *Ordered beta regression: A parsimonious, well-fitting model for continuous data with lower and upper bounds*. SocArXiv. https://osf.io/3r8a6/

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *37*(6), 1036–1073. https://doi.org/10.1111/cogs.12054

Lehmann, M. P., Xu, H. A., Liakoni, V., Herzog, M. H., Gerstner, W., & Preuschoff, K. (2019). One-shot learning and behavioral eligibility traces in sequential decision making. *elife*, *8*, Article e47463. https://doi.org/10.7554/eLife.47463

Lewis, D. (1974). Causation. *The Journal of Philosophy*, *70*(17), 556–567. https://doi.org/10.2307/2025310

Liljeholm, M. (2015). How multiple causes combine: Independence constraints on causal inference. *Frontiers in Psychology*, *6*, Article 1135. https://doi.org/10.3389/fpsyg.2015.01135

Liljeholm, M., & Cheng, P. W. (2009). The influence of virtual sample size on confidence and causal-strength judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 157–172. https://doi.org/10.1037/a0013972

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 intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332. https://doi.org/10.1016/j.cogpsych.2010.05.002

Lombrozo, T., & Vasilyeva, N. (2017). Causal explanation. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 415–432). Oxford University Press.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955–984. https://doi.org/10.1037/a0013256

Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, *122*(4), 700–734. https://doi.org/10.1037/a0039655

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, *37*, 205–220. https://doi.org/10.1146/annurev-neuro-071013-014017

Makowski, D., Ben-Shachar, M. S., Chen, S. A., & Lüdecke, D. (2019). Indices of effect existence and significance in the bayesian framework. *Frontiers in Psychology*, *10*, Article 2767. https://doi.org/10.3389/fpsyg.2019.02767

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Meyniel, F., & Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences*, *114*(19), E3859–E3868. https://doi.org/10.1073/pnas.1615773114

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as bayesian probability: From neural origins to behavior. *Neuron*, *88*(1), 78–92. https://doi.org/10.1016/j.neuron.2015.09.039

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PLOS ONE*, *14*(8), Article e0219704. https://doi.org/10.1371/journal.pone.0219704

Morris, A., Phillips, J. S., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). *Causal judgments approximate the effectiveness of future interventions*. PsyArxiv. https://osf.io/j8swe/

Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E., & Bahrami, B. (2017). The idiosyncratic nature of confidence. *Nature Human Behaviour*, *1*(11), 810–818. https://doi.org/10.1038/s41562-017-0215-1

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., Bello, P., Pearson, J., & De Brigard, F. (2022). Confidence and gradation in causal judgment. *Cognition*, *223*, Article 105036. https://doi.org/10.1016/j.cognition.2022.105036

O'Neill, K., Quillien, T., & Henne, P. (2022). *A counterfactual model of causal judgment in double prevention* [Conference session]. Conference in Computational Cognitive Neuroscience.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60. https://doi.org/10.1145/3241036

Perales, J. C., & Shanks, D. R. (2003). Normative and descriptive accounts of the influence of power and contingency on causal judgement. *The Quarterly Journal of Experimental Psychology Section A*, *56*(6), 977–1007. https://doi.org/10.1080/02724980244000738

Pescetelli, N., Rees, G., & Bahrami, B. (2016). The perceptual and social components of metacognition. *Journal of Experimental Psychology: General*, *145*(8), 949–965. https://doi.org/10.1037/xge0000180

Peters, M. A. (2022). Confidence in decision-making. *Oxford research encyclopedia of neuroscience*. Oxford University Press.

Peters, M. A., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E., & Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(7), Article 0139. https://doi.org/10.1038/s41562-017-0139

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences*, *23*(12), 1026–1040. https://doi.org/10.1016/j.tics.2019.09.007

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240

Quillien, T. (2020). When do we think that x caused y? *Cognition*, *205*, Article 104410. https://doi.org/10.1016/j.cognition.2020.104410

Quillien, T., & Barlev, M. (2021). *Causal judgment in the wild: Evidence from the 2020 us presidential election*. PsyArXiv. https://osf.io/r85tg/

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

Rasmussen, C., & Williams, C. (2005). *Gaussian processes for machine learning*. MIT Press.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266. https://doi.org/10.1038/nature08275

Riihimäki, J., & Vehtari, A. (2010). Gaussian processes with monotonicity information. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 645–652). Proceedings of Machine Learning Research.

Rouault, M., Weiss, A., Lee, J. K., Drugowitsch, J., Chambon, V., & Wyart, V. (2022). Controllability boosts neural and cognitive signatures of changes-of-mind in uncertain environments. *eLife*, *11*, Article e75038. https://doi.org/10.7554/eLife.75038

Samaha, J., & Denison, R. (2020). *The positive evidence bias in perceptual confidence is not post-decisional*. BioRxiv. https://doi.org/10.1101/2020.03.15.991513

Sartorio, C. (2020). More of a cause? *Journal of Applied Philosophy*, *37*(3), 346–363. https://doi.org/10.1111/japp.12370

Shanks, D. R. (1987). Acquisition functions in contingency judgment. *Learning and Motivation*, *18*(2), 147–166. https://doi.org/10.1016/0023-9690(87)90008-7

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193. https://doi.org/10.1016/j.tics.2014.01.006

Shekhar, M., & Rahnev, D. (2021). Sources of metacognitive inefficiency. *Trends in Cognitive Sciences*, *25*(1), 12–23. https://doi.org/10.1016/j.tics.2020.10.007

Solak, E., Murray-Smith, R., Leithead, W. E., Leith, D. J., & Rasmussen, C. E. (2003). *Derivative observations in gaussian process models of dynamic systems*. MIT Press.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, *126*(4), 323–348. https://doi.org/10.1037/0096-3445.126.4.323

Stan Development Team. (2020). *RStan: The R interface to Stan* (R package Version 2.21.2) [Computer software]. https://mc-stan.org/

Stan Development Team. (2021). *Stan modeling language users guide and reference manual* (Version 2.27) [Computer software]. https://mc-stan.org/docs/2_27/stan-users-guide/fit-gp-section.html#multiple-output-gaussian-processes

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020). Time and singular causation—A computational model. *Cognitive Science*, *44*(7), Article e12871. https://doi.org/10.1111/cogs.12871

Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science*, *10*(1), 242–257. https://doi.org/10.1111/tops.12309

Sytsma, J. (2021). Causation, responsibility, and typicality. *Review of Philosophy and Psychology*, *12*(4), 699–719. https://doi.org/10.1007/s13164-020-00498-2

Sytsma, J. (2022). The responsibility account. In P. Willemsen & A. Wiegmann (Eds.), *Advances in experimental philosophy of causation* (pp. 145–164). Bloomsbury Publishing.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems* (pp. 59–65). MIT Press.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. https://doi.org/10.1007/BF00122574

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science*, *42*(4), 1265–1296. https://doi.org/10.1111/cogs.12605

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637. https://doi.org/10.1111/cogs.12101

Weiss, A., Chambon, V., Lee, J. K., Drugowitsch, J., & Wyart, V. (2021). Interacting with volatile environments stabilizes hidden-state inference and its brain signatures. *Nature Communications*, *12*(1), Article 2228. https://doi.org/10.1038/s41467-021-22396-6

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111. https://doi.org/10.1037/0096-3445.136.1.82

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*(1), 1–50. https://doi.org/10.1215/00318108-2005-001

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, Article 21111. https://doi.org/10.3389/fnins.2012.00001

Received June 6, 2023
Revision received April 22, 2024
Accepted May 2, 2024 ■