Measuring and Modeling Confidence in Human Causal Judgment

Kevin O'Neill
Center for Cognitive Neuroscience
Duke University
kevin.oneill@duke.edu

Paul Henne
Department of Philosophy
Lake Forest College
phenne@mx.lakeforest.edu

John Pearson
Center for Cognitive Neuroscience
Duke University
john.pearson@duke.edu

Felipe De Brigard
Center for Cognitive Neuroscience
Duke University
felipe.debrigard@duke.edu

Abstract

The human capacity for causal judgment has long been thought to depend on an ability to consider counterfactual alternatives: the lightning strike caused the forest fire because had it not struck, the forest fire would not have ensued. To accommodate psychological effects on causal judgment, a range of recent accounts of causal judgment have proposed that people probabilistically sample counterfactual alternatives from which they compute a graded index of causal strength. While such models have had success in describing the influence of probability on causal judgments, among other effects, we show that these models make further untested predictions: probability should also influence people's metacognitive confidence in their causal judgments. In a large (N=3020) sample of participants in a causal judgment task, we found evidence that normality indeed influences people's confidence in their causal judgments and that these influences were predicted by a counterfactual sampling model. We take this result as supporting evidence for existing Bayesian accounts of causal judgment.

Keywords: causal judgment; metacognition; counterfactual thinking

Introduction

Judgments about cause and effect are thought to be central to the way people decide who or what is responsible for an outcome (Chockler & Halpern, 2004; Knobe & Fraser, 2008; Malle, Guglielmo, & Monroe, 2014) or explain how a particular state affairs came to be (Lombrozo, 2007; Lombrozo & Vasilyeva, 2017). In machine learning, causal judgment is considered a major requirement for systems that generate robust predictions in a range of circumstances and intervene in the world, and much recent work accordingly focuses on how to develop systems capable of representing, learning, and making use of causal information (Dasgupta et al., 2019; Gershman, 2017; Gershman, Norman, & Niv, 2015; Pearl, 2019). Drawing on both of these literatures, computational models of human causal judgment seek to explain why people tend to think of some events as more causal than other events, while also providing a tractable framework for implementing such judgments in artificial agents. Among the many possibilities, counterfactual sampling models have had particular success (Cheng, 1997; Cheng & Novick, 1990; Icard, Kominsky, & Knobe, 2017; Quillien, 2020; Spellman, 1997). These models account for known effects of probability (Gerstenberg & Icard, 2020; Henne, O'Neill, Bello, Khemlani, & De Brigard, 2021; Icard et al., 2017; Knobe & Fraser, 2008), the presence of alternative causes (Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015; Lagnado, Gerstenberg, & Zultan, 2013), temporal recency (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Henne, Kulesza, Perez, & Houcek, 2021; Spellman, 1997), and foreseeability (Kirfel & Lagnado, 2021) on causal judgments, among other phenomena. Counterfactual sampling models have even been shown to predict eye movements during causal judgment (Bello, Lovett, Briggs, & O'Neill, 2018; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017) and judgments of omissive causation (Gerstenberg & Stephan, 2021; Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019).

However, while there is a vast amount of research on causal judgment, little is known about how and whether people are able to evaluate the accuracy and reliability of their causal judgments (but see Liljeholm, 2015, 2020; Liljeholm & Cheng, 2009). In this paper, taking ideas from models of metacognition in perception and decision-making (Ma & Jazayeri, 2014; Meyniel & Dehaene, 2017; Meyniel, Sigman, & Mainen, 2015; Pouget, Drugowitsch, & Kepecs, 2016), we propose the first computational model (to our knowledge) of metacognitive confidence in human causal judgments, or simply *causal metacognition*. Comparing several variations of this model to participants' ratings, we found that one of these variations was able to simultaneously predict mean causal judgment and mean confidence in a simple causal judgment task. In the Discussion, we argue that our results constitute strong evidence in favor of this model and we discuss implications for future research.

Counterfactual sampling and causal judgment

Before extending the predictions of counterfactual sampling models to the domain of causal metacognition, we will first briefly review how they account for causal judgments themselves. Counterfactual sampling models assume that people encode causal relationships between variables using a causal graph consisting of exogenous variables $\mathcal{U}$ whose causes are not explicitly modeled, endogenous variables $\mathcal{V}$ which are determined as a function of the exogenous variables $\mathcal{U}$, and a set of structural equations $\mathcal{F}$ that encode the dependence of $\mathcal{V}$ on $\mathcal{U}$ (represented as edges in the graph). Here we will focus on the causal structure depicted in Figure 1, known as an unshielded collider (Pearl, 2019). In this structure, an effect E is produced by two generative causes: a focal cause C and an alternate cause A. That is, $\mathcal{U} = \{U_C, U_A\}$, $\mathcal{V} = \{C, A, E\}$. We focus on two versions of this structure

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

for the case of binary variables. In the *conjunctive* structure, both causes are necessary for the effect to occur (i.e., $\mathcal{F} = \{C = U_C,\ A = U_A,\ E = \min(C,A)\}$). In the *disjunctive* structure, either cause is individually sufficient to produce the effect (i.e., $\mathcal{F} = \{C = U_C,\ A = U_A,\ E = \max(C,A)\}$).

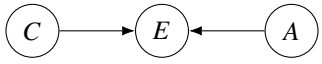

Figure 1: A causal graph depicting the relationships between an effect E as produced by a focal cause C and an alternate cause A.

Counterfactual sampling models aim to predict people's causal judgments of the extent to which $C = c$ caused $E = e$ given the above causal graph and the observations $C = c$, $A = a$, and $E = e$. To do so, they propose that people sample alternative possibilities according to the internal model:

$$\begin{aligned} C' &\sim \text{Bernoulli}(\theta_C) \\ A' &\sim \text{Bernoulli}(\theta_A) \\ \kappa_{C\to E} &= f(C', A', \mathcal{F}) \end{aligned}$$

where $\theta_C \propto P(C)$, $\theta_A \propto P(A)$. The value $\kappa_{C\to E}$ corresponds to some measure of the difference (or contribution) made by C to E for each sampled possibility, where the function f determines exactly how the difference made by C to E is quantified (Table 1). In addition to two models that were originally formulated using the sampling algorithm above (Icard et al., 2017; Quillien, 2020), we also include three classic measures of causal strength that can be estimated under the same algorithm, though it is important to note that these measures were not originally derived with this particular problem, algorithm, or causal structure in mind (Cheng, 1997; Cheng & Novick, 1990; Spellman, 1997). Following Morris, Phillips, Gerstenberg, and Cushman (2019), our goal is not to evaluate these models in their original context, but rather to test whether the measures they provide (construed as quantifications of difference-making for single events) predict causal judgments in this domain.

For instance, the ΔP model uses a measure that corresponds to the difference between the value that E would have taken if $C = 1$ (denoted $E_{C=1,A=A'}$) and the value it would have taken if $C = 0$ ($E_{C=0,A=A'}$) (Cheng & Novick, 1990). The Power PC model uses the same metric as ΔP but with a different normalization (Cheng, 1997). The crediting causality model (Spellman, 1997) is also similar to the ΔP model, but it uses the average value of the effect overall, and not the value of the effect when the cause is absent, as baseline. More recently, the necessity-sufficiency model computes the impact of C by determining whether it was sufficient for E (if $C' = 1$) or whether it was necessary for E (if $C' = 0$) (Icard et al., 2017). Finally, in our causal structure of interest the counterfactual effect size model is equivalent to ΔP except that it uses a normalization based on the standard deviations $\sigma_{C'}$ and $\sigma_{E'}$ of C' and E', respectively (Quillien, 2020).

Given a choice of f, counterfactual sampling generates a probability distribution $P(\kappa_{C\to E})$, which corresponds to the belief that $C = c$ caused $E = e$. These models typically assume that causal judgments are reports of the expected causal strength $\mathbb{E}[\kappa_{C\to E}]$. This summary creates natural interpretations for many choices of f. For instance, ΔP reduces to the average causal effect of C on E, and the counterfactual effect size model is simply the correlation between C' and $E_{C=C',A=A'}$ in the sampled possibilities (Cheng & Novick, 1990; Quillien, 2020). Since each of the above models has seen empirical support, we will extend each of them to predict confidence in causal judgments.

Counterfactual sampling and metacognition

While research on causal judgment has typically focused on $\mathbb{E}[\kappa_{C\to E}]$, the expected causal strength of C on E, counterfactual sampling models of causal judgment assume that people have access to samples from the distribution $P(\kappa_{C\to E})$. In the domains of perception and decision-making, recent models of metacognition based on Bayesian decision theory have suggested that the information provided by the distribution over a decision variable is sufficient (if not necessary) to produce metacognitive assessments of confidence (Ma & Jazayeri, 2014; Meyniel & Dehaene, 2017; Meyniel et al., 2015; Navajas et al., 2017; Pouget et al., 2016; Yeung & Summerfield, 2012). For binary decisions (e.g. whether a stimulus is present or absent), this distribution allows one to compute the probability that the decision is correct as a measure of confidence (Fleming & Daw, 2017; Hangya, Sanders, & Kepecs, 2016; Kepecs, Uchida, Zariwala, & Mainen, 2008; Kiani & Shadlen, 2009). However, even in contexts where all of the relevant variables are binary, causal judgments are thought to be continuous or graded such that an event can be seen as more or less causal (Danks, 2017; Halpern & Hitchcock, 2015; O'Neill, Henne, Bello, Pearson, & De Brigard, 2021). Thankfully, a number of options exist for quantifying uncertainty in continuous decisions: the variance, the standard deviation, the coefficient of variation, and the entropy are all natural candidates for modeling people's reports of confidence in their causal judgments (Liljeholm, 2015; Meyniel et al., 2015). Conceptually, the variance, standard deviation, and coefficient of variation all propose that people are less confident in their causal judgments if their belief $P(\kappa_{C\to E})$ is imprecise or variable, though each measures variability on a slightly different scale. Similarly, entropy proposes that people are more confident in their causal judgments if $P(\kappa_{C\to E})$ carries more information. Table 2 summarizes each of these measures and provides their formulae for the case where $\kappa_{C\to E}$ is Bernoulli-distributed, which for the causal structures of interest applies to all of the models in Table 1 except for the counterfactual effect size model, in which case the only difference is that $\text{Var}(\kappa_{C\to E}) = \frac{\text{Var}(C')}{\text{Var}(E')}\mathbb{E}[\kappa_{C\to E}](1-\mathbb{E}[\kappa_{C\to E}])$.

Thus, our model of causal metacognition is a simple conjunction of counterfactual sampling models of causal judgment and Bayesian models of metacognition: causal judg-

Table 1: Causal strength metrics from five counterfactual sampling models

Model	$\kappa_{C\to E}$
ΔP (Cheng & Novick, 1990)	$\Delta P(C',\, A',\, \mathcal{F}) = E_{C=1,A=A'} - E_{C=0,A=A'}$
Power PC (Cheng, 1997)	$\mathrm{PPC}(C',\, A',\, \mathcal{F}) = \frac{\Delta P(C',A',\mathcal{F})}{1-E_{C=0,A=A'}}$
Crediting Causality (Spellman, 1997)	$\mathrm{CC}(C',\, A',\, \mathcal{F}) = E_{C=1,A=A'} - E_{C=C',A=A'}$
Necessity-Sufficiency (Icard et al., 2017)	$\mathrm{NS}(C',\, A',\, \mathcal{F}) = C' * E_{C=1,A=A'} + (1-C') * (1-E_{C=0,A=A'})$
Counterfactual Effect Size (Quillien, 2020)	$\mathrm{CES}(C',\, A',\, \mathcal{F}) = \frac{E_{C=1-C',A=A'} - E_{C=C',A=A'}}{1-2C'} \frac{\sigma_{C'}}{\sigma_{E'}}$

Table 2: Four confidence metrics for counterfactual sampling models

Measure	
Variance	$\mathrm{Var}(\kappa_{C\to E}) = \mathbb{E}[\kappa_{C\to E}](1-\mathbb{E}[\kappa_{C\to E}])$
SD	$\sigma_{\kappa_{C\to E}} = \sqrt{\mathrm{Var}(\kappa_{C\to E})}$
CV	$\mathrm{CV}(\kappa_{C\to E}) = \sigma_{\kappa_{C\to E}} / \mathbb{E}[\kappa_{C\to E}]$
Entropy	$\mathrm{H}(\kappa_{C\to E}) = -\Sigma P(\kappa_{C\to E}) \log(P(\kappa_{C\to E}))$

ments are reports of the expected difference the cause made to the effect (i.e., $\mathbb{E}[\kappa_{c\to E}]$) and confidence ratings are reports of the expected certainty in this estimate (e.g., inversely related to $\sigma_{\kappa_{c\to E}}$). To test this model, we replicated and extended a recent study measuring quantitative shifts in causal judgments with respect to the probabilities of the focal 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 (Icard et al., 2017; Kominsky et al., 2015; Morris et al., 2019). Each of the above measures of uncertainty predict that people's confidence in their causal judgments should also vary with $P(C)$ and $P(A)$. Accordingly, we also measure participants' confidence in their causal judgments.

Methods

Participants

3020 participants were recruited 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 IRB. Participants completed the task in an average of 7.5 minutes and were compensated $0.75. 118 (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. Data were analyzed from the remaining 2902 participants (mean age = 36.93, standard deviation age = 13.23, 49% female).

Materials

Stimuli were six vignettes similar to the vignette used in Morris et al. (2019). 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 and only 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, participants were shown an image that briefly summarized the vignette and also defined the probability of each candidate cause. All materials and code are accessible via the Open Science Framework. For example, participants were shown the following vignette along with the image in Figure 2:

> A person, Joe, played a casino game where he reached into two boxes and blindly drew a ball from each box. 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. 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?**

Procedure

In a $10 \times 10 \times 2 \times 6$ within-participants design (probability of focal cause: $\{.1, .2, \ldots 1\}$; probability of alternate cause: $\{.1, .2, \ldots 1\}$; causal structure: Conjunctive/Disjunctive; vignette), participants read one version of each of the six vignettes. The probability of each candidate cause and the causal structure were randomly assigned for each vignette. The probability of each candidate cause could take any value between .1 and 1 with increments of .1, and the order of vignettes was randomized. For each vignette, participants read

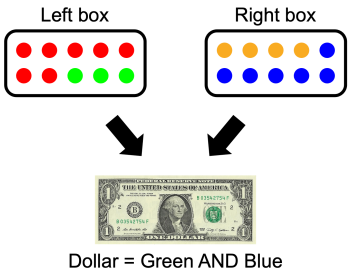

Figure 2: Example stimulus. In this example, a character wins a dollar if and only if they draw a green ball from the left box (with probability .3) and they draw a blue ball from the right box (with probability .6).

the vignette and inspected an image which added information about the probability of each event. On the same screen, participants responded to the questions "To what degree did [the outcome occur] because [the focal cause occurred]?" and "How confident are you in your response to the previous question?" on 1000-point continuous slider scales ranging from "not at all" (coded as 0) to "totally" (coded as 1).

Analysis

To determine the effects of the probability of the focal 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). We estimated mean causal judgment and mean confidence as the inferred mean from separate GPs for conjunctive and disjunctive causal structures. Importantly, using a GP model allowed us to account for known non-linear effects of probability on causal judgments in a way that maximizes statistical power (Morris et al., 2019) and to account for known correlations between mean confidence, mean causal judgment, and variability in causal judgments (O'Neill et al., 2021). To test for changes in causal judgments and confidence ratings with respect to the probability of each cause, we also jointly estimated the gradients of each GP (Riihimäki & Vehtari, 2010; Solak, Murray-Smith, Leithead, Leith, & Rasmussen, 2003). All GPs were modeled on a latent logit scale with an Ordered Beta likelihood (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. Full specification of the model and prior distributions is available via the Open Science Framework. We considered any parameter with a 95% highest density posterior interval excluding zero as statistically significant.

Results

Causal Judgment

We first sought to replicate previous results showing that causal judgments vary as a function of the probability of the focal cause (i.e., the cause that we ask participants to judge)

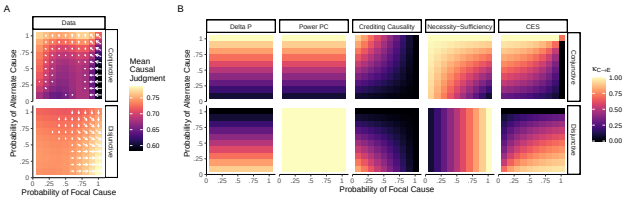

Figure 3: Inferred mean causal judgment (A) compared to model predictions (B). Arrows indicate significant gradients in mean causal judgment with respect to the probability of the focal or alternate causes. Color scales differ between the data and the model predictions to better illustrate trends.

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). Figures 3A and 3B depict mean causal judgment and predictions from each model, respectively. In conjunctive structures, causal judgments of the focal cause C tended to decrease with the probability of the focal cause and increase with the probability of the alternate cause. In disjunctive structures, we found the opposite result: causal judgments tended to increase with the probability of the focal cause and decrease with the probability of the alternate cause. The white arrows in Figure 3 indicate regions where these trends were significant.

We then asked whether these patterns in causal judgments were predicted by counterfactual sampling models. To answer this question, we computed correlations between inferred mean causal judgment and the predictions from each model. Figure 5 (left panel) depicts the performance of each model along this metric. As found in previous work (Morris et al., 2019; Quillien, 2020), we found that counterfactual sampling models were largely successful in predicting causal judgments. In particular, the counterfactual effect size model had the highest correlation with mean causal judgment for both conjunctive ($r = .88$, 95% $HDI = [.81, .93]$) and disjunctive ($r = .74$, 95% $HDI = [.50, .93]$) causal structures. All models significantly predicted causal judgments in conjunctive structures, and all models except the Power PC ($r = 0$) and Crediting Causality ($r = -.04$, 95% $HDI = [-.37, .29]$) models significantly predicted causal judgments in disjunctive structures.

Confidence

Next, we asked whether people's confidence in their causal judgments also varied with respect to the probability of the focal and alternate causes. Figure 4 depicts mean confidence in causal judgments alongside predictions from each model. Because model predictions were naturally on the scale of uncertainty (with larger numbers indicating less certainty), we normalized all model predictions to the range [0, 1], with 0 indicating uncertainty and 1 indicating certainty. In conjunctive structures, people tended to be more confident in their causal judgments as the probability of the focal cause decreased and

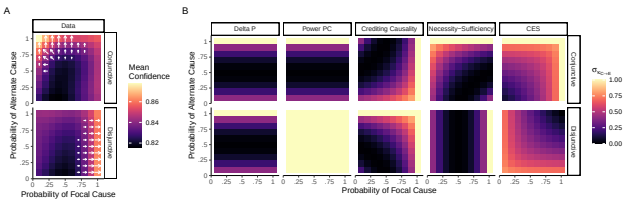

Figure 4: Mean confidence in causal judgment (A) compared to model predictions using the standard deviation of predictions of causal judgments (B). Arrows indicate significant gradients in mean confidence with respect to the probability of the focal or alternate causes. For visibility, color scales differ between the data and the model predictions, and model predictions were normalized to the range [0, 1], with 0 indicating uncertainty and 1 indicating certainty.

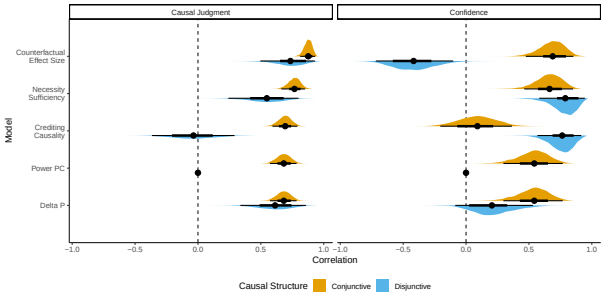

Figure 5: Model performance for causal judgments and confidence ratings in conjunctive (orange) and disjunctive (blue) causal structures. While most models perform well at predicting causal judgments, only the Necessity-Sufficiency model predicts causal judgments and confidence for both causal structures. Points indicate posterior medians, thick error bars indicate 66% highest density intervals, and thin error bars indicate 95% highest density intervals.

as the probability of the alternate cause increased. In contrast, in disjunctive structures, people tended to be more confident as the probability of the focal cause increased. White arrows in Figure 4 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 to the corresponding effects on causal judgment. As such, the confidence judgments may have been subject to a ceiling effect, limiting the generalizability of these findings.

Finally, we tested whether Bayesian models of metacognition, in conjunction with counterfactual sampling models of causal judgment, predicted participants' confidence in their causal judgments. As with causal judgments, we correlated the inferred mean confidence with the predictions from each model. For simplicity, we depict results only using the standard deviation $\sigma_{\kappa_{C\to E}}$. Results were qualitatively similar using other metrics, which can be found via the Open Science Framework. Figure 5 (right panel) depicts the performance of each model along this metric. While the counterfactual effect size model again performed the best in conjunctive structures ($r = .69$, 95% $HDI = [.48, .85]$), it performed the worst in disjunctive structures ($r = -.42$, 95% $HDI = [-72., -.12]$). Other models were significantly able to predict confidence in either conjunctive structures or disjunctive structures, but the only model to significantly predict confidence in *both* conjunctive ($r = .77$, 95% $HDI = [.66, .85]$) and disjunctive ($r = .66$, 95% $HDI = [.46, .85]$) structures was the necessity-sufficiency model.

Discussion

In this article, we proposed an extension of counterfactual sampling models of causal judgment to additionally model participants' confidence in their causal judgments. Our extension, following recent work in metacognition, is simple: whereas people report causal judgments as the expected causal strength $\mathbb{E}[\kappa_{C\to E}]$, they report confidence as the uncertainty in this estimate, using e.g. the standard deviation $\sigma_{\kappa_{C\to E}}$. This extension of counterfactual sampling models made the novel prediction that people should be more or less confident in their causal judgments depending on the probability of each of the contributing causes of the effect. However, different variations of the model differed in exactly how confidence should change: some predicted confidence should increase with the probability of the focal cause, others predicted that it should depend only on the probability of the alternate cause, some predicted that confidence would be a nonlinear function of the two probabilities, and still others predicted no changes in confidence whatsoever.

To test the different variations of our model, we replicated and extended an experiment by Morris et al. (2019) which demonstrated that causal judgments tended to decrease with the probability of the focal cause and increase with the probability of the alternate cause in conjunctive causal structures (i.e., when both causes are individually necessary for the effect), but they tended to increase with the probability of the focal cause and decrease with the probability of the alternate cause in disjunctive causal structures (i.e., when either cause is individually sufficient for the effect). Our experiment reproduced these results, and most variations of the model were able to significantly predict causal judgments in both causal structures (Morris et al., 2019; Quillien, 2020).

Extending these findings, we also measured the degree to which participants were confident in their causal judgments. As with causal judgments, we found that participants' confidence decreased with the probability of the focal cause and increased with the probability of the alternate cause in conjunctive causal structures, but their confidence increased with the probability of the focal cause in disjunctive causal structures. These patterns were only significantly predicted by a single version of the model: the necessity-sufficiency model (Icard et al., 2017). Because each measure was developed solely to explain causal judgments (with no regard for con-

fidence), testing their metacognitive predictions provides an especially strong test of the generalizability of these models. In this sense, it is not surprising that most models were unable to predict the observed changes in confidence. In contrast, we take the ability of the necessity-sufficiency model to account for such changes as a clear sign of its predictive utility.

However, more work is needed to investigate how people make metacognitive assessments of their 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. They accordingly reported very high confidence overall. But people most often make causal judgments in the presence of these types of uncertainty, in addition to mere probabilistic uncertainty. In addition, people usually obtain information relevant for causal judgment from a range of sources and modalities which may vary in their degrees of credibility. In contrast, in our study, participants were provided full information from a single reliable source. Relaxing this assumption may help in determining how and when people update causal judgments and how this updating affects their confidence. Future work should also explore the ways in which metacognitive assessments of causal judgments impact subsequent cognition, particularly in relation to real-world domains like elections (Quillien & Barlev, 2021) where outcomes have a significant and lasting impact. It is widely known that metacognition of perceptual and value-based decisions affects learning, exploration, and changes of mind (Folke, Jacobsen, Fleming, & De Martino, 2016; Kepecs et al., 2008; Shea et al., 2014). We would expect causal metacognition to have similar effects on behavior.

Finally, future work may explore alternative mechanisms for confidence in causal judgments. Our model of causal metacognition is a *first-order* model in that both causal judgments and confidence ratings emerge from a distribution over the same underlying variable $\kappa_{C\rightarrow E}$ (Fleming & Daw, 2017). Causal metacognition, however, may be better modeled as a *second-order* phenomenon whereby causal judgments and confidence arise from separate decision variables. Alternatively, confidence in causal judgments may come from a more heuristic approach (Adler & Ma, 2018). In addition to deepening our understanding of the human ability to track confidence in causal judgments, adjudicating between these different architectures may provide crucial insights toward the development of metacognitive artificial agents.

In sum, we proposed an extension of counterfactual sampling models of human causal judgment to additionally predict confidence in those judgments. When compared to 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), e1006572.

Bello, P., Lovett, A. M., Briggs, G., & O'Neill, K. (2018). An attention-driven computational model of human causal reasoning. In *Proceedings of the 40th annual meeting of the cognitive science society.*

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.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1), 1–32.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological review*, *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.

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

Danks, D. (2017). Singular causation. *The oxford handbook of causal reasoning*, 201–215.

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., ... Kurth-Nelson, Z. (2019). Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, *124*(1), 91.

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), 1–8.

Gershman, S. J. (2017). Reinforcement learning and causal models. *The Oxford handbook of causal reasoning*, 295.

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43–50.

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

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

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, *216*, 104842.

Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of*

Science, *66*(2), 413–457.

Hangya, B., Sanders, J. I., & Kepecs, A. (2016). A mathematical framework for statistical decision confidence. *Neural Computation*, *28*(9), 1840–1858.

Henne, P., Kulesza, A., Perez, K., & Houcek, A. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition*, *212*, 104708.

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

Henne, P., O'Neill, K., Bello, P., Khemlani, S., & De Brigard, F. (2021). Norms affect prospective causal judgments. *Cognitive Science*, *45*(1), e12931.

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

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.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *science*, *324*(5928), 759–764.

Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, 104721.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral psychology*, *2*, 441–8.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Kubinec, R. (2020). Ordered beta regression: A parsimonious, well-fitting model for continuous data with lower and upper bounds. *SocArXiv. March*, *2*.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive science*, *37*(6), 1036–1073.

Liljeholm, M. (2015). How multiple causes combine: independence constraints on causal inference. *Frontiers in psychology*, *6*, 1135.

Liljeholm, M. (2020). Neural correlates of causal confounding. *Journal of cognitive neuroscience*, *32*(2), 301–314.

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.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive psychology*, *55*(3), 232–257.

Lombrozo, T., & Vasilyeva, N. (2017). Causal explanation. *Oxford handbook of causal reasoning*, 415–432.

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual review of neuroscience*, *37*, 205–220.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186.

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.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as bayesian probability: From neural origins to behavior. *Neuron*, *88*(1), 78–92.

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

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.

O'Neill, K., Henne, P., Bello, P., Pearson, J., & De Brigard, F. (2021). Degrading causation. *OSF Preprints*.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience*, *19*(3), 366.

Quillien, T. (2020). When do we think that x caused y? *Cognition*, *205*, 104410.

Quillien, T., & Barlev, M. (2021). Causal judgment in the wild: evidence from the 2020 us presidential election. *PsyArXiv*.

Riihimäki, J., & Vehtari, A. (2010). Gaussian processes with monotonicity information. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 645–652).

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.

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.

Stan Development Team. (2020). *RStan: the R interface to Stan.* Retrieved from `http://mc-stan.org/` (R package version 2.21.2)

Stan Development Team. (2021). *Stan modeling language users guide and reference manual, version 2.27.* Retrieved from `https://mc-stan.org/docs/2_27/stan-users-guide/fit-gp-section.html#multiple-output-gaussian-processes`

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.