Measuring and Modeling Confidence in Human Causal Judgment

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

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

John Pearson
Center for Cognitive Neuroscience
Department of Biostatistics & Bioinformatics
Department of Psychology and Neuroscience
Department of Electrical & Computer Engineering
Duke University
john.pearson@duke.edu

Felipe De Brigard
Center for Cognitive Neuroscience
Department of Psychology and Neuroscience
Department of Philosopy
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.

1 Introduction

Judgments about cause and effect are thought to be central to the way people decide who or what is responsible for an outcome [1, 2, 3] or explain how a particular state affairs came to be [4, 5]. 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

Presented at the Workshop on Metacognition in the Age of AI: Challenges and Opportunities, 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

accordingly focuses on how to develop systems capable of representing, learning, and making use of causal information [6, 7, 8, 9]. 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 [10, 11, 12, 13, 14]. These models account for known effects of probability [2, 13, 15, 16], the presence of alternative causes [17, 18], temporal recency [12, 19, 20], and foreseeability [21] on causal judgments, among other phenomena. Counterfactual sampling models have even been shown to predict eye movements during causal judgment [22, 23] and judgments of omissive causation [24, 25].

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 [26, 27, 28]). In this paper, taking ideas from models of metacognition in perception and decision-making [29, 30, 31, 32], 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 this ability to predict both causal judgments and confidence constitutes strong evidence in favor of this model and we discuss implications for future research on causal judgment, metacognition, and artificial agents.

1.1 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 [9]. In this structure, an effect E is produced by two causes: a focal cause C and an alternate cause A. That is, $\mathcal{U} = \{C, A\}$, $\mathcal{V} = \{E\}$. We focus on two versions of this structure for the case of binary variables. In the *conjunctive* structure, both causes are necessary for the effect to occur (i.e., $\mathcal{F} = \{E = \min(C, A)\}$). In the *disjunctive* structure, either cause is individually sufficient to produce the effect (i.e., $\mathcal{F} = \{E = \max(C, A)\}$).

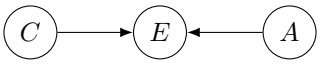


Figure 1: A causal graph depicting the relationships between an effect E as produced by a focal cause C and an alternate cause A.

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:

$$C' \sim \text{Bernoulli}(\theta_C) \qquad A' \sim \text{Bernoulli}(\theta_A)$$
$$\mathcal{U}' = \{C = C', A = A'\} \qquad \kappa_{C \to E} = f(\mathcal{U}', \mathcal{F})$$

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. Because different variants of counterfactual sampling models differ in how they quantify $\kappa_{C \to E}$, each version of the model uses a unique specification of the function f outlined in Table 1.

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 $C = 0$ ($E_{C=0,A=A'}$) [10]. The Power PC model uses the same metric as ΔP but with a different normalization [11]. The crediting causality model [12] 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 computing

Table 1: Causal strength metrics from five counterfactual sampling models

Model	$\kappa_{C\rightarrow E}$
ΔP [10]	$\Delta P(\mathcal{U}', \mathcal{F}) = E_{C=1,A=A'} - E_{C=0,A=A'}$
Power PC [11]	$\text{PPC}(\mathcal{U}', \mathcal{F}) = \frac{\Delta P(\mathcal{U}', \mathcal{F})}{1-E_{C=0,A=A'}}$
Crediting Causality [12]	$\text{CC}(\mathcal{U}', \mathcal{F}) = E_{C=1,A=A'} - E_{C=C',A=A'}$
Necessity-Sufficiency [13]	$\text{NS}(\mathcal{U}', \mathcal{F}) = C' * E_{C=1,A=A'} + (1 - C') * (1 - E_{C=0,A=A'})$
Counterfactual Effect Size [14]	$\text{CES}(\mathcal{U}', \mathcal{F}) = \frac{E_{C=1-C',A=A'} - E_{C=C',A=A'}}{1-2C'} \frac{\sigma_{C'}}{\sigma_{E'}}$

whether it was sufficient for E (if $C' = 1$) or whether it was necessary for e' (if $C' = 0$) [13]. 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 [14].

Given a choice of f, counterfactual sampling generates a probability distribution $P(\kappa_{C\rightarrow E})$, 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 $\mathbb{E}[\kappa_{C\rightarrow E}]$. 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 [10, 14]. Since each of the above models has seen empirical support, we will extend each of them to predict confidence in causal judgments.

1.2 Counterfactual sampling and metacognition

While research on causal judgment has typically focused on $\mathbb{E}[\kappa_{C\rightarrow E}]$, the expected causal strength of C on E, counterfactual sampling models of causal judgment assume that people have access to the full distribution $P(\kappa_{C\rightarrow E})$. 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 full probability distribution over a decision variable is sufficient (if not necessary) to produce metacognitive assessments of confidence [29, 30, 32, 33, 31, 34]. 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 [35, 36, 37, 38]. 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 [39, 40, 41]. 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 [30, 27]. Conceptually, the variance, standard deviation, and coefficient of variation all propose that people are more confident in their causal judgments if their belief $P(\kappa_{C\rightarrow E})$ 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 $P(\kappa_{C\rightarrow E})$ carries more information. Table 2 summarizes each of these measures and provides their formulae for the case where $\kappa_{C\rightarrow E}$ 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 $\text{Var}(\kappa_{C\rightarrow E}) = \frac{\text{Var}(C')}{\text{Var}(E')}\mathbb{E}[\kappa_{C\rightarrow E}](1 - \mathbb{E}[\kappa_{C\rightarrow E}])$.

Thus, our model of causal metacognition is a simple conjunction of counterfactual sampling models of causal judgment and Bayesian models of metacognition: causal judgments are reports of the expected difference the cause made to the effect (i.e., $\mathbb{E}[\kappa_{c\rightarrow E}]$) and confidence ratings are reports of the certainty in this estimate (e.g., inversely related to $\sigma_{\kappa_{c\rightarrow E}}$). 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 $P(A)$ [42]. 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

Table 2: Four confidence metrics for counterfactual sampling models

Measure	
Variance	$\text{Var}(\kappa_{C\to E}) = \mathbb{E}[\kappa_{C\to E}](1 - \mathbb{E}[\kappa_{C\to E}])$
Standard Deviation	$\sigma_{\kappa_{C\to E}} = \sqrt{\text{Var}(\kappa_{C\to E})}$
Coefficient of Variation	$\text{CV}(\kappa_{C\to E}) = \sigma_{\kappa_{C\to E}}/\mathbb{E}[\kappa_{C\to E}]$
Entropy	$\text{H}(\kappa_{C\to E}) = -\Sigma P(\kappa_{C\to E})\ \log(P(\kappa_{C\to E}))$

structures [18, 13, 42]. 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.

2 Methods

2.1 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).

2.2 Materials

Stimuli were six vignettes similar to the vignette used in [42]. 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 stimuli are available in Appendix A and 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?**

2.3 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

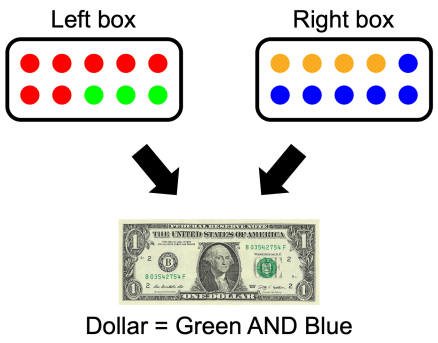


Figure 2: Example stimulus. In this example, a character wins a dollar if and only if they draw a green ball from the left box (with probability .3) and they draw a blue ball from the right box (with probability .6).

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 the vignette and inspected its corresponding image which added information about the probability of each event occurring. 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 continuous slider scales ranging from "not at all" (coded as 0) to "totally" (coded as 1).

2.4 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 [43, 44, 45]. We estimated mean causal judgment and mean confidence as the inferred mean from separate GPs for conjunctive and disjunctive causal structures. 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 [46, 47]. All GPs were modeled on a latent logit scale with an Ordered Beta likelihood [48], 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 in Appendix C. We considered any parameter with a 95% highest density posterior interval excluding zero as statistically significant.

3 Results

3.1 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) and the alternate cause (i.e., the cause that participants do not judge) [18, 13, 42]. 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 [42, 14], 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 ΔP ($r = 0$) and Crediting Causality ($r = -.04$, 95% $HDI = [-.37, .29]$) models significantly predicted causal judgments in disjunctive structures.

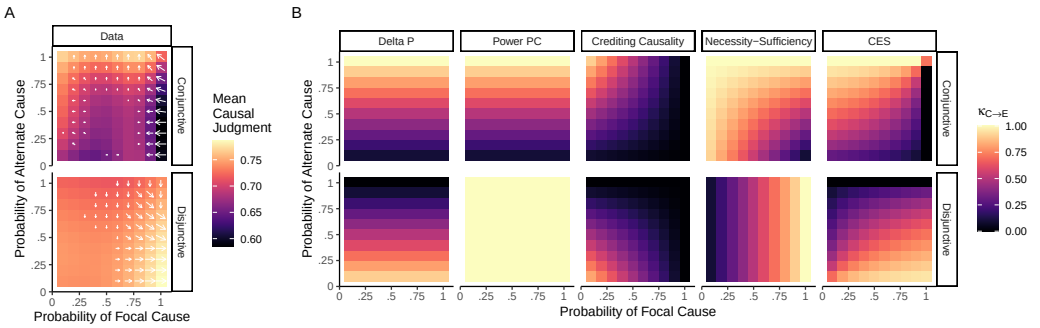


Figure 3: Inferred mean causal judgment (A) compared to model predictions (B). Arrows indicate significant gradients in mean causal judgment with respect to the probability of the focal or alternate causes. Color scales differ between the data and the model predictions to better illustrate trends.

3.2 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 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.

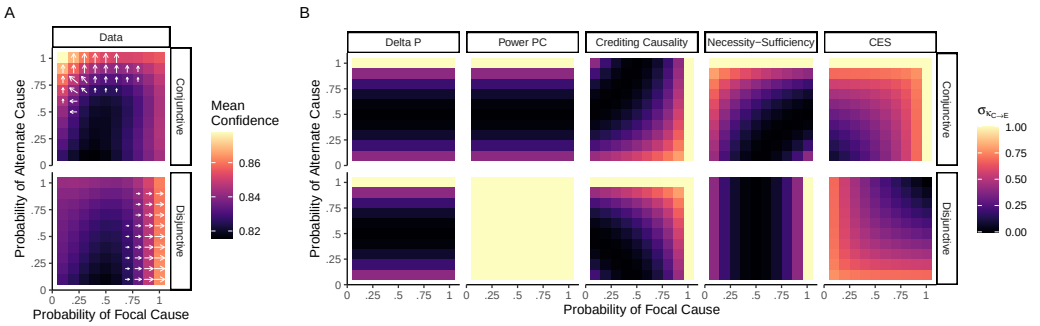


Figure 4: Mean confidence in causal judgment (A) compared to model predictions using the standard deviation of predictions of causal judgments (B). Arrows indicate significant gradients in mean confidence with respect to the probability of the focal or alternate causes. For visibility, color scales differ 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.

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

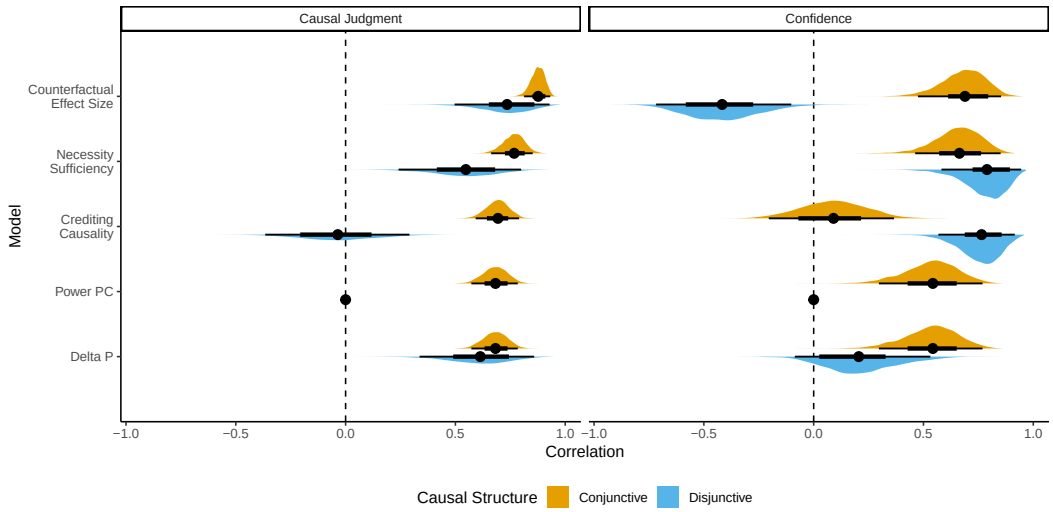


Figure 5: Model performance for causal judgments and confidence ratings in conjunctive (red) and disjunctive (blue) causal structures. While most models perform well at predicting causal judgments, only the Necessity-Sufficiency model predicts causal judgments and confidence for both causal structures.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

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 in Appendix B. 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.

4 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 models 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. [42], 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 of the counterfactual sampling models were able to significantly predict causal judgments in both causal structures [42, 14].

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 [13]. Because each measure of causal strength was developed solely to explain causal judgments (with no regard for confidence), 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 [49] 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 [37, 50, 51]. 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 $\kappa_{C \to E}$ [35]. 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 [52]. 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 [13]. 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.

Acknowledgments and Disclosure of Funding

This research was supported by the Office of Naval Research grant N00014-17-1-2603 to FDB. We would also like to thank Benjamin Eva and Elika Bergelson for their helpful discussion.

References

[1] Hana Chockler and Joseph Y Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

[2] Joshua Knobe and Ben Fraser. Causal judgment and moral judgment: Two experiments. *Moral psychology*, 2:441–8, 2008.

[3] Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. A theory of blame. *Psychological Inquiry*, 25(2):147–186, 2014.

[4] Tania Lombrozo. Simplicity and probability in causal explanation. *Cognitive psychology*, 55 (3):232–257, 2007.

[5] Tania Lombrozo and Nadya Vasilyeva. Causal explanation. *Oxford handbook of causal reasoning*, pages 415–432, 2017.

[6] Samuel J Gershman, Kenneth A Norman, and Yael Niv. Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, 5:43–50, 2015.

[7] Samuel J Gershman. Reinforcement learning and causal models. *The Oxford handbook of causal reasoning*, page 295, 2017.

[8] Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019.

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

[10] Patricia W Cheng and Laura R Novick. A probabilistic contrast model of causal induction. *Journal of personality and social psychology*, 58(4):545, 1990.

[11] Patricia W Cheng. From covariation to causation: A causal power theory. *Psychological review*, 104(2):367, 1997.

[12] Barbara A Spellman. Crediting causality. *Journal of Experimental Psychology: General*, 126 (4):323, 1997.

[13] Thomas F Icard, Jonathan F Kominsky, and Joshua Knobe. Normality and actual causal strength. *Cognition*, 161:80–93, 2017.

[14] Tadeg Quillien. When do we think that x caused y? *Cognition*, 205:104410, 2020.

[15] Tobias Gerstenberg and Thomas Icard. Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, 149(3):599, 2020.

[16] Paul Henne, Kevin O'Neill, Paul Bello, Sangeet Khemlani, and Felipe De Brigard. Norms affect prospective causal judgments. *Cognitive Science*, 45(1):e12931, 2021.

[17] David A Lagnado, Tobias Gerstenberg, and Ro'i Zultan. Causal responsibility and counterfactuals. *Cognitive science*, 37(6):1036–1073, 2013.

[18] Jonathan F Kominsky, Jonathan Phillips, Tobias Gerstenberg, David Lagnado, and Joshua Knobe. Causal superseding. *Cognition*, 137:196–209, 2015.

[19] Neil R Bramley, Tobias Gerstenberg, Ralf Mayrhofer, and David A Lagnado. Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(12):1880, 2018.

[20] Paul Henne, Aleksandra Kulesza, Karla Perez, and Augustana Houcek. Counterfactual thinking and recency effects in causal judgment. *Cognition*, 212:104708, 2021.

[21] Lara Kirfel and David Lagnado. Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, 212:104721, 2021.

[22] Tobias Gerstenberg, Matthew F Peterson, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. Eye-tracking causality. *Psychological science*, 28(12):1731–1744, 2017.

[23] Paul Bello, Andrew M Lovett, Gordon Briggs, and Kevin O'Neill. An attention-driven computational model of human causal reasoning. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, 2018.

[24] Paul Henne, Laura Niemi, Ángel Pinillos, Felipe De Brigard, and Joshua Knobe. A counterfactual explanation for the action effect in causal judgment. *Cognition*, 190:157–164, 2019.

[25] Tobias Gerstenberg and Simon Stephan. A counterfactual simulation model of causation by omission. *Cognition*, 216:104842, 2021.

[26] Mimi Liljeholm and Patricia W Cheng. The influence of virtual sample size on confidence and causal-strength judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(1):157, 2009.

[27] Mimi Liljeholm. How multiple causes combine: independence constraints on causal inference. *Frontiers in psychology*, 6:1135, 2015.

[28] Mimi Liljeholm. Neural correlates of causal confounding. *Journal of cognitive neuroscience*, 32(2):301–314, 2020.

[29] Wei Ji Ma and Mehrdad Jazayeri. Neural coding of uncertainty and probability. *Annual review of neuroscience*, 37:205–220, 2014.

[30] Florent Meyniel, Mariano Sigman, and Zachary F Mainen. Confidence as bayesian probability: From neural origins to behavior. *Neuron*, 88(1):78–92, 2015.

[31] Alexandre Pouget, Jan Drugowitsch, and Adam Kepecs. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience*, 19(3):366, 2016.

[32] Florent Meyniel and Stanislas Dehaene. Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences*, 114(19):E3859–E3868, 2017.

[33] Joaquin Navajas, Chandni Hindocha, Hebah Foda, Mehdi Keramati, Peter E Latham, and Bahador Bahrami. The idiosyncratic nature of confidence. *Nature human behaviour*, 1(11): 810–818, 2017.

[34] Nick Yeung and Christopher Summerfield. Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1310–1321, 2012.

[35] Stephen M Fleming and Nathaniel D Daw. Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, 124(1):91, 2017.

[36] Balázs Hangya, Joshua I Sanders, and Adam Kepecs. A mathematical framework for statistical decision confidence. *Neural Computation*, 28(9):1840–1858, 2016.

[37] Adam Kepecs, Naoshige Uchida, Hatim A Zariwala, and Zachary F Mainen. Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455(7210):227–231, 2008.

[38] Roozbeh Kiani and Michael N Shadlen. Representation of confidence associated with a decision by neurons in the parietal cortex. *science*, 324(5928):759–764, 2009.

[39] David Danks. Singular causation. *The oxford handbook of causal reasoning*, pages 201–215, 2017.

[40] Joseph Y Halpern and Christopher Hitchcock. Graded causation and defaults. *The British Journal for the Philosophy of Science*, 66(2):413–457, 2015.

[41] Kevin O'Neill, Paul Henne, Paul Bello, John Pearson, and Felipe De Brigard. Degrading causation. *OSF Preprints*, 2021.

[42] Adam Morris, Jonathan Phillips, Tobias Gerstenberg, and Fiery Cushman. Quantitative causal selection patterns in token causation. *PloS one*, 14(8):e0219704, 2019.

[43] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1):1–32, 2017.

[44] Stan Development Team. RStan: the R interface to Stan, 2020. URL http://mc-stan.org/. R package version 2.21.2.

[45] Stan Development Team. Stan modeling language users guide and reference manual, version 2.27, 2021. URL https://mc-stan.org/docs/2_27/stan-users-guide/fit-gp-section.html#multiple-output-gaussian-processes.

[46] Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 645–652. JMLR Workshop and Conference Proceedings, 2010.

[47] Ercan Solak, Roderick Murray-Smith, William E Leithead, Douglas J Leith, and Carl Edward Rasmussen. Derivative observations in gaussian process models of dynamic systems. *MIT Press*, 2003.

[48] Robert Kubinec. Ordered beta regression: A parsimonious, well-fitting model for continuous data with lower and upper bounds. *SocArXiv. March*, 2, 2020.

[49] Tadeg Quillien and Michael Barlev. Causal judgment in the wild: evidence from the 2020 us presidential election. *PsyArXiv*, 2021.

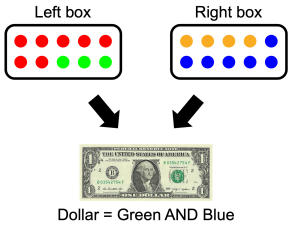
[50] Nicholas Shea, Annika Boldt, Dan Bang, Nick Yeung, Cecilia Heyes, and Chris D Frith. Supra-personal cognitive control and metacognition. *Trends in cognitive sciences*, 18(4):186–193, 2014.

[51] Tomas Folke, Catrine Jacobsen, Stephen M Fleming, and Benedetto De Martino. Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1(1):1–8, 2016.

[52] William T Adler and Wei Ji Ma. Comparing bayesian and non-bayesian accounts of human confidence reports. *PLoS computational biology*, 14(11):e1006572, 2018.

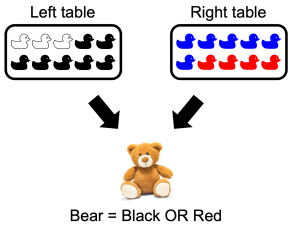
A Vignettes

A.1 Casino



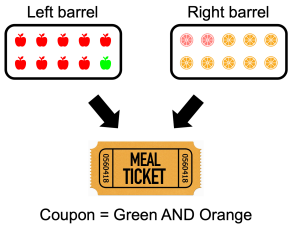
A person, Joe, played a casino game where he reached into two boxes and blindly drew a ball from each box.	
Conjunctive: In this game, he wins a dollar if and only if he gets a green ball from the left box and a blue ball from the right box. If he doesn't get a green ball from the left box or he doesn't 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?**	

A.2 Ducks



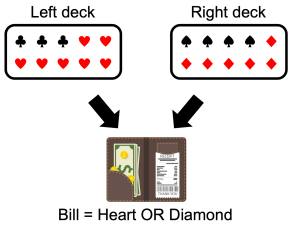
Frida played a carnival game where she selected rubber ducks with duck stickers on the bottom from two separate tables.	
Conjunctive: In this game, she wins a stuffed animal if and only if the duck she chooses from the left table has a black sticker on the bottom and the duck she chooses from the right table has a red sticker on the bottom. If the duck she chooses from the left table has a white sticker on the bottom or the duck she chooses from the right table has a blue sticker on the bottom, then she does not win the stuffed animal.	**Disjunctive:** In this game, she wins a stuffed animal if the duck she chooses from the left table has a black sticker on the bottom or the duck she chooses from the right table has a red sticker on the bottom (or both). If the duck she chooses from the left table has a white sticker on the bottom and the duck she chooses from the right table has a blue sticker on the bottom, then she does not win the stuffed animal.
Frida grabbed a rubber duck from each table with each hand, and picked them up at the same time. When she flipped them over, she saw that the duck from the left table had a black sticker on the bottom and the duck from the right table had a red sticker on the bottom. So, Frida won the stuffed animal.	
To what degree did Frida win the stuffed animal because she chose a duck with a black sticker on the bottom from the left table? **How confident are you in your response to the previous question?**	

A.3 Fruit



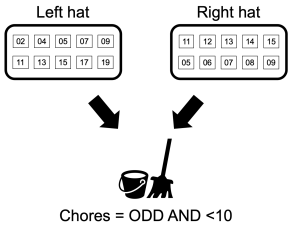
Isaac and June played a game at a birthday party where they blindfolded themselves and bobbed for fruit from two separate barrels.	
Conjunctive: In this game, they win a coupon for a restaurant if and only if Isaac gets a green apple from the left barrel and June gets a lemon from the right barrel. If Isaac gets a red apple from the left barrel or if June gets a lime from the right barrel, they don't win the coupon.	**Disjunctive:** In this game, they win a coupon for a restaurant if Isaac gets a green apple from the left barrel or June gets a lemon from the right barrel (or both). If Isaac gets a red apple from the left barrel and June gets a lime from the right barrel, they don't win the coupon.
On the count of three, Isaac and June each pulled a fruit from their respective barrels. After removing their blindfolds, they discovered that Isaac got a green apple from the left barrel and June got a lemon from the right barrel. So, they won the restaurant coupon.	
To what degree did Isaac and June win the restaurant coupon because Isaac pulled a green apple from the left barrel? **How confident are you in your response to the previous question?**	

A.4 Cards



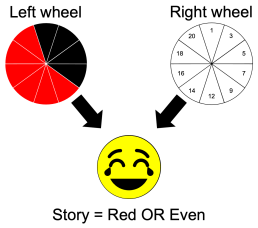
Rebecca and her friends came up with a bet to pay for drinks last night by drawing cards from two decks with 10 cards in each deck.	
Conjunctive: According to this bet, Rebecca has to pay for everyone's drinks if and only if she picks a card of hearts from the left deck and a card of diamonds from the right deck. If she picks a card of clubs from the left deck or a card of spades from the right deck, then Rebecca doesn't need to pay for anyone's drinks.	**Disjunctive:** According to this bet, Rebecca has to pay for everyone's drinks if she picks a card of hearts from the left deck or if she picks a card of diamonds from the right deck (or both). If she picks a card of clubs from the left deck and a card of spades from the right deck, then Rebecca doesn't need to pay for anyone's drinks.
Rebecca blindly picked a card from each deck. Her friend flipped them over at the same time to reveal that Rebecca picked a card of hearts from the left deck and a card of diamonds from the right deck. So, Rebecca paid for everyone's drinks that night.	
To what degree did Rebecca pay for everyone's drinks because she picked a card of hearts from the left deck? **How confident are you in your response to the previous question?**	

A.5 Chores



Victor and his roommates agreed on a plan to assign chores for the week by drawing numbers out of two hats.	
Conjunctive: According to this plan, Victor has to do the chores this week if and only if he draws an odd number from the hat on the left and he draws a number less than 10 from the hat on the right. If Victor draws an even number from the hat on the left or a number greater than 10 from the hat on the right, then he doesn't have to do the chores this week.	**Disjunctive:** According to this plan, Victor has to do the chores this week if he draws an odd number from the hat on the left or he draws a number less than 10 from the hat on the right (or both). If Victor draws an even number from the hat on the left and a number greater than 10 from the hat on the right, then he doesn't have to do the chores this week.
Victor reached a hand into each hat and, at the same time, he drew an odd number from the hat on the left and a number less than 10 from the hat on the right. So, Victor did the chores this week.	
To what degree did Victor do the chores this week because he drew an odd number from the hat on the left? **How confident are you in your response to the previous question?**	

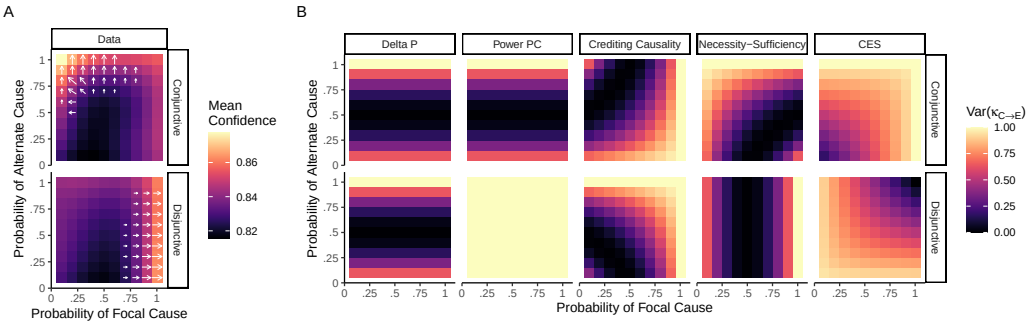
A.6 Party



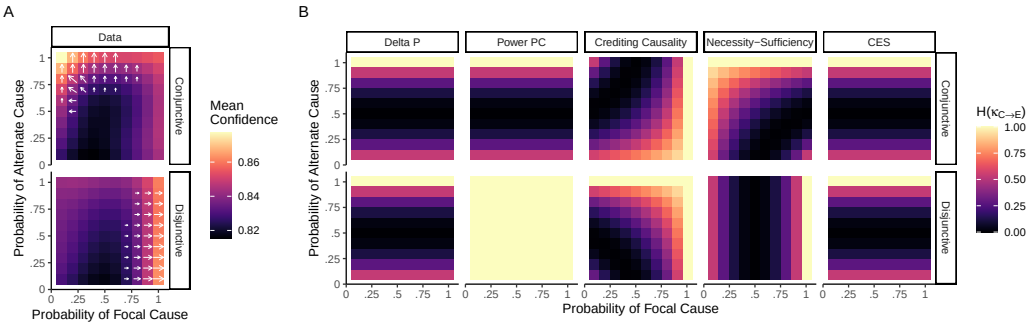
Beth played a party game with some friends last night where they spun two different wheels.	
Conjunctive: In this game, she must reveal an embarrassing story about herself if and only if the wheel on the left lands on a red area and the wheel on the right lands on an even number. If the wheel on the left lands on black or the wheel on the right lands on an odd number, then she doesn't have to tell her friends an embarrassing story.	**Disjunctive:** In this game, she must reveal an embarrassing story about herself if the wheel on the left lands on a red area or the wheel on the right lands on an even number (or both). If the wheel on the left lands on black and the wheel on the right lands on an odd number, then she doesn't have to tell her friends an embarrassing story.
Beth grabbed each wheel with a separate hand and spun them both. At the same time, the wheel on the left landed on red and the wheel on the right landed on an even number. So, Beth told her friends an embarrassing story from her past.	
To what degree did Beth tell her friends an embarrassing story about herself because the wheel on the left landed on red? **How confident are you in your response to the previous question?**	

B Alternative measures of confidence

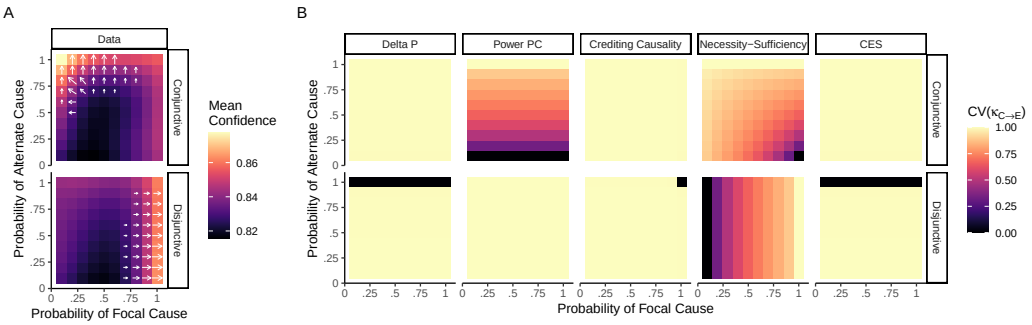
B.1 Variance



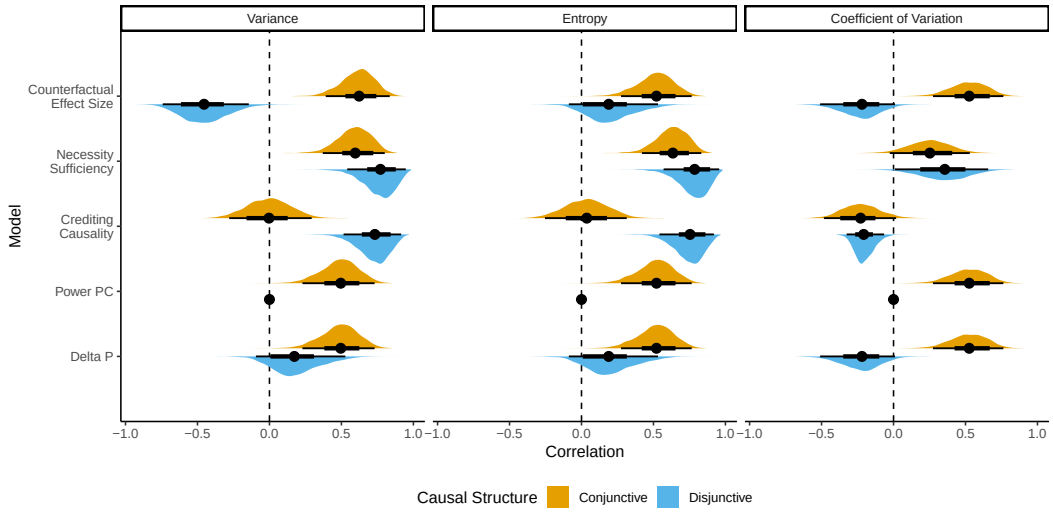
B.2 Entropy



B.3 Coefficient of variation



B.4 Model comparisons



C Gaussian Process Model

Let $y_{i,r} \in [0,1]$ be the response on trial $i = 1 \ldots N$ of response r (causal judgment or confidence rating). We wish to model these ratings as a function of both the probability of the focal and alternate causes ($P(C)$ and $P(A)$, respectively). We assume that ratings for different values of focal and alternate causes are related to each other by a latent Gaussian Process (f_s, $s = 1, 2$) modeled separately for the conjunctive and disjunctive causal structures on the logit scale. More specifically, we assume

$$y_{i,r} \sim \text{OrderedBeta}(\text{logit}^{-1}(f_{s[i],\mathbf{p}[i],r}), \phi_r, \theta_{0,r}, \theta_{1,r}), \tag{1}$$

where $s[i] \in \{\text{conjunctive}, \text{disjunctive}\}$ and $\mathbf{p}[i]$ represent the causal structure and $(P(C), P(A))$ values of observation i, respectively, and the Ordered Beta distribution [48] is defined by

$$\text{OrderedBeta}(\mu, \phi, \theta_0, \theta_1) = \begin{cases} 0, & \mu < \theta_0 \\ \text{Beta}(\mu\phi, (1-\mu)\phi) & \theta_0 \leq \mu \leq \theta_1 \\ 1, & \mu > \theta_1 \end{cases} \tag{2}$$

where μ is the mean, ϕ is a precision parameter, and θ_0 and θ_1 are thresholds that force the predicted response to 0 or 1, respectively. For the precision and threshold parameters of the Ordered Beta distribution, we assume the following weakly-informative priors:

$$\begin{aligned} \phi_r &\sim \text{HalfNormal}(0, 1) \\ \theta_{1,r} - \theta_{0,r} &\sim \text{HalfNormal}(0, 3) \end{aligned}$$

Finally, we assume that the latent mean responses are determined by a Gaussian Process ($f_{s,r}$, multivariate in r, independent in s) with kernel $K_{s,rr'}(\mathbf{p}, \mathbf{p}')$, where $\mathbf{p} = (P(C), P(A))$. We factorize this as a product over two kernels, scaled by the marginal standard deviation α_s [45]:

$$K_{s,rr'}(\mathbf{p}, \mathbf{p}') = \alpha_s \; k_s(\mathbf{p}, \mathbf{p}') \; \Omega_{s,rr'} \tag{3}$$

$$k_s(\mathbf{p}, \mathbf{p}') = e^{-0.5 \sum_d (p_d - p'_d)^2 / \rho^2_{s,d}}. \tag{4}$$

Here, the squared exponential kernel $k_s(\mathbf{p}, \mathbf{p}')$ defines the covariance between the values of f_s at different values of $\mathbf{p}$, while $\Omega_{rr'}$ defines the covariance between the two dimensions of f_s (i.e., causal judgments and confidence ratings). To efficiently sample f_s with the kernel $K_{s,rr'}(\mathbf{p}, \mathbf{p}')$, we further decompose $k_s(\mathbf{p}, \mathbf{p}')$ and $\Omega_{s,rr'}$ using the Cholesky decomposition:

$$k_s(\mathbf{p}, \mathbf{p}') = L_{k_s(\mathbf{p},\mathbf{p}')} L^T_{k_s(\mathbf{p},\mathbf{p}')} \tag{5}$$

$$\Omega_{s,rr'} = L_{\Omega_{s,rr'}} L^T_{\Omega_{s,rr'}}. \tag{6}$$

Whitening f_s with $L_{k_s(\mathbf{p},\mathbf{p}')}$ and $L_{\Omega_{s,rr'}}$ yields i.i.d. samples from the standard normal distribution. As a result, we can sample f_s as the product

$$f_s = L_{k_s(\mathbf{p},\mathbf{p}')} \; \eta \; L_{\Omega_{s,rr'}} \qquad\qquad \eta_{ij} \sim \mathcal{N}(0, 1). \tag{7}$$

This sampling procedure yields the desired result that $f_{s,r} \sim GP(\mathbf{0}, K_{s,rr'}(\mathbf{p}, \mathbf{p}'))$.

Finally, for the hyperparameters $\rho_{s,r}$ (defining the length-scale of $k_s(\mathbf{p}, \mathbf{p}')$) and $\alpha_{s,r}$ (defining the marginal standard deviation), we use the following weakly-informative priors:

$$\rho_{s,r} \sim \text{InvGamma}(4, 0.5) \tag{8}$$

$$\alpha_{s,r} \sim \mathcal{N}(0, 2). \tag{9}$$