Variability in Causal Judgments

Ivar Kolvoort (i.r.kolvoort@uva.nl)
University of Amsterdam, Faculty of Social and Behavioural Sciences

Zachary J. Davis (zach.davis@stanford.edu)
Stanford University, Department of Psychology

Leendert Van Maanen (l.vanmaanen@uu.nl)
Utrecht University, Social and Behavioral Sciences

Bob Rehder (bob.rehder@nyu.edu)
New York University, Department of Psychology

Abstract

Causal knowledge permeates higher-level cognition, influencing processes from judgment and decision-making to categorization. Yet little is known about the mental processes by which people come to make causal judgments. In this paper, we study the *within-participant variability* of judgments to gain insight into the processes used to generate them. We establish that causal judgments exhibit substantial within-participant variability. This variability varies by inference type and is related to the extent to which participants commit Markov violations. The consistency and systematicity of this variability must be explained by any complete model of how people make causal judgments. The systematic study of both within- and between-person variability broadens the scope of behavior that can be studied in causal cognition and promotes the evaluation of formal models of the underlying process. We provide a benchmark data set that can be used for such model evaluation.

Keywords: causal reasoning; process models; variability; Markov violations

Introduction

Causal relationships are a central way in which humans experience the world. Causal knowledge affects what decisions we make, how we categorize objects, and what counts as a good explanation (for summary, see Sloman, 2005). One of the main tools in studying causal cognition has been the theoretical framework known as causal graphical models (CGMs) (Pearl, 2000). CGMs have been shown to provide a generally good account of the causal judgments that people make. However, causal graphical models provide a computational level account that specifies *what* causal judgements are made, but not necessarily *how* people make them. Given the importance of causal knowledge to higher-level cognition, surprisingly little attention has been given to the *processes* by which people make such sophisticated judgments. In addition, recent empirical investigations have identified multiple systematic deviations from CGM predictions in human data (Rehder, 2014; Rehder & Waldmann, 2017; Davis & Rehder, 2020; Rottman & Hastie, 2016; Davis & Rehder, 2017). To account for these deviations, researchers have developed multiple, mostly descriptive, theories (Rehder, 2014, 2018; Rottman & Hastie, 2016; Trueblood et al., 2017). These theories have been hard to distinguish as they have been developed to account for the same data, and they vary in how much light they shed on the process by which people generate causal judgments.

How can we evaluate which process generated a judgment? The predominant approach is to assess the predictions of multiple models against the average judgments of participants. This approach is principled and effective, but in a field as rich as causal cognition, utilizing only averaged data has not been able to convincingly identify the best model out of the multitude that have been proposed (Rehder, 2014, 2018; Rottman & Hastie, 2016). Other data can help with this underdetermination problem. For example, in judgment and decision making the popular diffusion decision model has exhibited considerable success in not merely accounting for mean judgments, but also explaining full distributions of response variables (Ratcliff et al., 2016). In this project, we use the full distribution of causal judgments as a new source of information about underlying cognitive processes involved.

A few studies have remarked on the considerable variability in human causal judgments (Davis & Rehder, 2020; Rehder, 2014; Rottman & Hastie, 2016). However, it is hitherto unclear to what extent that variability represents within- or between-participant variability. The standard method to discriminate between these sources of variability, and to do participant-level distributional analyses, is to run experiments employing a repeated measures design. Such measures have not yet been used in the study of causal reasoning due to practical concerns.

A major difficulty in conducting a repeated measures design for causal judgments is the likelihood that judgments are not independent. Other areas that commonly use repeated measures often have stimuli such as random-dot motion arrays, which can be presented repeatedly without participants' awareness. Causal judgments do not have this property. Stimuli like ours that are composed of discrete symbols (such as states of causal variables) are susceptible to be recognized and memorized. This can be a problem particularly for studying higher order cognition, such as causal reasoning, due to its more deliberative and conscious nature. In fact, storing previous judgments for future use has been proposed to be an important source of computational savings for limited agents (Dasgupta & Gershman, 2021). Our challenge was to design an experiment that elicits independent judgments for repeated causal queries by reducing the likelihood that participants' judgments are informed by prior computations or memory.

Here we present the first study of probabilistic causal reasoning using a repeated measures design to elicit multiple in-

dependent judgments. The experiment makes use of particular inferences and knowledge domains to jump the methodological hurdles just described. We use a symmetrical causal structure, query participants regarding both the absence and presence of causal factors, and use the same parameterization across different domains in order to obtain repeated measures.

As this is the first study of within-participant variability in causal judgments we aim to answer some foundational questions. Firstly, we aim to establish whether there is meaningful within-participant variability in causal judgments. Secondly, we look to compare variability across different inference types. Are there differences between forward (from cause to effect) and backward (from effect to cause) inferences? Does the information on which a participant is to base their inference impact variability? Thirdly, we investigate whether individual level variability is related to a tendency to commit an important systematic reasoning error known as Markov violations. We then describe potential models of variability in the causal reasoning process and provide a comparison of the observed variability against their qualitative predictions. We conclude by discussing the connections between the patterns of variation observed in our study with existing findings in causal cognition and opportunities for the use of full response distributions in the study of how people reason with causal information.

Experiment

Materials. We tested causal judgments about variables in five domains: biology, astronomy, economics, meteorology, and sociology. Participants were first told that the domain they were about to study included three binary variables. For example, in the domain of economics they were told that interest rates could be either low or normal, trade deficits that were small or normal, and retirement savings that were high or normal.

Participants were then presented with a description of two causal relations that formed a common cause network in which one variable (henceforth referred to as Y) was a cause of the two others (X_1 and X_2). Each causal relationship was generative and included a description of the mechanism responsible for that relationship. An example in the domain of economics is "Low interest rates cause small trade deficits. The low cost of borrowing money leads businesses to invest in the latest manufacturing technologies, and the resulting low-cost products are exported around the world." All these materials have been validated by and used in multiple other studies (Rehder, 2014, 2018; Rehder & Waldmann, 2017).

Procedure. Subjects first studied several screens of information about the overall task that established the domains being studied and the types of inferences that would be presented during the study. Then, for each domain, initial screens presented a cover story and a description of the domain's three variables and subsequent screens presented the two causal links and a diagram of those links. A common cause network was used in every domain, and participants were informed that each variable's baserate was 50% and that each cause produced its effect "75% of the time".

When ready, participants were asked three multiple-choice questions to assess their understanding of the causal relationships. This comprehension check established that they had learned which variables were causally related, the direction of those relationships, and that the relationships were probabilistic rather than deterministic. Participants were given three attempts to answer all questions correctly. Once they answered all questions correctly or after the third attempt participants could continue with the experiment.

Subjects were then presented with the inference test. Each trial presented the values of one or two variables and asked to predict the state of another. For example, a subject might be told that an economy has low interest rates and a normal trade deficit and be asked the probability of it having a high level of retirement savings. Subjects entered their response by moving a tick on a rating scale whose ends were labeled 0% and 100%. As an attention check, participants were asked a comprehension check question at the end of each block. The order of the five domains, and the 24 test questions within each domain, was randomized for each participant.

Design and Participants. We chose six particular inference types to be tested based on the relevant comparisons they would allow. Firstly we wanted to compare diagnostic or 'backward' inferences in which one has to judge the probability of a cause based on knowledge of its effects with non-diagnostic (or 'forward' or 'predictive') inferences in which one reasons from cause to effect. Second, we assessed the effect of the information on which participants had to condition their inference: consistent information (e.g. $X_i = 1, Y = 1$), inconsistent information (e.g. $X_i = 1, Y = 0$), and incomplete information (e.g. $X = 1$ and Y unknown). These factors lead to the six inference types presented in Table 1. To obtain repeated measurements, within each domain each inference type was queried four times by (a) varying whether the role of X_i was filled by X_1 or X_2 (possible because of the symmetry of the common cause structure) and (b) asking about both the presence and the absence of the to-be-inferred variable (using $P(X_i = 1|Y) = 1 - P(X_i = 0|Y)$). This resulted in each inference type being queried 20 times over the five domains and a total of 120 queries per participant. Table 1 also presents the normative conditional probabilities based on the aforementioned base rates (50%) and causal strengths (75%).

It is noteworthy that all of the non-diagnostic inferences have the same normative probability of 80%. These inferences have been shown to exhibit "Markov violations", a systematic pattern of responses in which, rather than adhering to the independence between certain causal variables stipulated by CGM theory, participants' responses are instead influenced by the value of the independent and hence irrelevant variable (Rehder, 2014; Rottman & Hastie, 2016). For these inferences, the value of one effect (X_j) should not provide information regarding the other effect (X_i) once the value of Y is known.

Table 1: Inference Types and Normative Answers

	Diagnostic	Non-diagnostic
Consistent	$P(Y\|X_i = 1, X_j = 1)$ = 94%	$P(X_i\|Y = 1, X_j = 1)$ = 80%
Incomplete	$P(Y\|X_i = 1)$ = 80%	$P(X_i\|Y = 1)$ = 80%
Inconsistent	$P(Y\|X_i = 1, X_j = 0)$ = 50%	$P(X_i\|Y = 1, X_j = 0)$ = 80%

All participants made all judgments for all five domains. 37 participants were recruited from Prolific (www.prolific.co) and received £5.70 for on average 47 minutes ($SD = 20.1$) of participation. 8 (22%) participants were removed from analyses for failing at least two attention checks, as had been established by the authors before the running of the study.

Results

Figure 1 illustrates the overall response distributions per inference type. The first aspect to note is that the distributions vary by judgment type. If the only source of variability is unrelated to the process by which causal judgments are generated (such as general response noise), we would expect similar variability across judgments. The bimodality of the response distributions in Figure 1 is also noteworthy. In particular, we observe a "spike" of responses at 50%, which has been reported previously (Rottman & Hastie, 2016). This peak at 50% seems to vary along the Information factor, with the largest peaks for inconsistent inferences and smallest for inferences with consistent information. As expected, the peak is largest for inconsistent diagnostic inferences for which the normative answer is 50%.

Figure 2 shows the means of within-participant standard deviations and mean judgments per inference type. Note in Figure 2 that while variability differs by inference type, it does not track with the mean, suggesting that these results are not driven by an artifact of the scoring system. We tested whether variability differs over the inference types using an repeated measures ANOVA with the standard deviation in responses as dependent variables and Diagnostic (yes, no) and Information (consistent, incomplete, inconsistent) as factors. We find that the main effects of both Diagnostic ($F(1,140) = 32.0,\ p < .001,\ BF > 100$) and Information ($F(2,140) = 11.7,\ p < .001,\ BF > 100$), as well as their interaction ($F(2,140) = 6.70,\ p = .002,\ BF = 18.6$) are significant. These results indicate that there was more within-participant variability for Diagnostic inferences compared to Non-diagnostic ones and (surprisingly) more for Consistent inferences compared to Incomplete inferences. These differences over judgment types suggest that variability results from some underlying process of generating causal judgments.

To test whether variability and Markov violations are related, we first separated participants into three (low, medium, high) equally sized groups based on the standard deviation of their responses on non-diagnostic inferences. Figure 3 plots the mean non-diagnostic judgments by variability group, revealing an apparent increase in non-normative responding as variability increases. We conducted an repeated measures ANOVA on the non-diagnostic responses with Information and the variability grouping factor as factors. We find a significant main effect of Information ($F(2,1546) = 39.4,\ p < .001,\ BF > 100$), which indicates that overall participants committed Markov violations, as normatively the Information factor should not have an effect as the normative response to all non-diagnostic inferences is 80%. We find no evidence for a main effect of the variability grouping variable ($F(2,26) = 2.11,\ p = .14,\ BF = 0.323$), indicating that participants with more variable judgments do not give smaller or larger estimates for the non-diagnostic inferences. Most interestingly, we find very strong evidence for an interaction between Information and the grouping variable ($F(4,1546) = 6.62,\ p < .001,\ BF > 100$), indicating that high variability participants commit larger Markov violations. This interaction is illustrated in Figure 3 by the thick black line, which becomes steeper (larger Markov violations) for the higher variability groups.

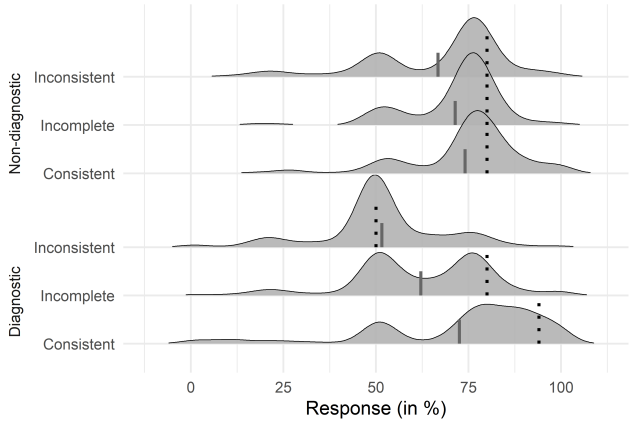

Figure 1: Overall response distributions per inference type. Vertical grey lines indicate mean responses. Dotted vertical black lines indicate normative response

Sources of variability

What processes explain the variability in responses to causal queries? As a guide for future research, in this section we outline a number of candidate models of the variability in conditional probability judgments. While fitting these models against participants' response distributions is beyond the scope of this paper, we discuss the correspondence of their qualitative predictions with the results of our experiment.

One possibility is that the observed variability in responses is entirely independent from the cognitive process by which a causal judgment is generated. It could be that people have a stable causal representation and strategy to arrive at a causal judgment, but that the process of responding to a query re-

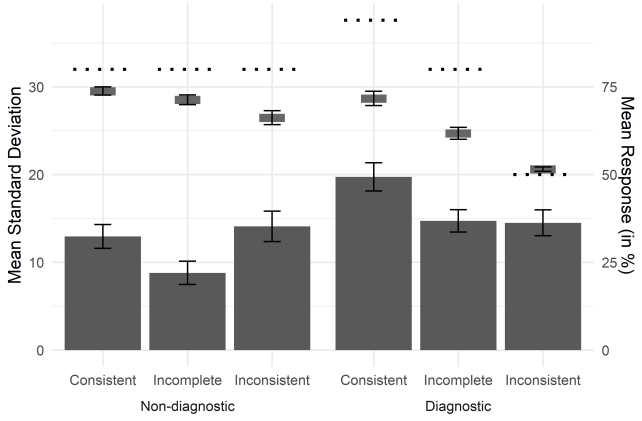

Figure 2: Barplot: Mean within-participant standard deviations per inference type. Floating dashes: Mean responses per inference type. Black vertical lines indicate standard error. Horizontal dotted lines indicate normative probability.

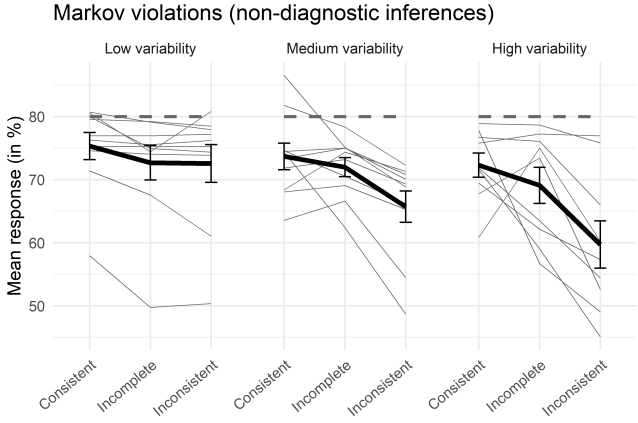

Figure 3: Plots of markov violation per variability group. Participants were first separated into three (low, medium, high) equally sized groups based on the variability in their responses on non-diagnostic inferences. Grey thin lines represent the mean responses of individual participants on the non-diagnostic inferences. Black thick lines represent the mean responses per group, the vertical bars indicate standard errors. The dashed lines represent the normative response of 80%.

sults in some noise, e.g. through motor noise in using a slider or some general task noise. In this case one would expect response distributions that are centered at the normative answer, such as predicted by the Beta inference model (Rottman & Hastie, 2016). Our findings provide evidence against this possibility: response distributions are often multi-modal (see Figure 1), and variability differs by inference type and seems to be related to patterns of non-normative responding.

Another possibility could be that the source of variability in causal judgments stems from uncertainty about the parameters of the described causal network. For example, rather than believing that the causal strength of A on B is precisely .75, this value may have some variance. Because the CGM framework models causal judgments as being computed from a causal network, this would result in variation in the resultant causal judgments. Such an account may explain increased variability in diagnostic inferences, as according to the CGM framework these require the processing of an additional parameter, the base rate of the cause (Fernbach et al., 2011). It is unclear how this approach would explain why judgments where two pieces of information are given are more variable than only one piece, as the CGM framework would predict that there is no change in the number of parameters that need to be considered. In addition, this CGM-based account is incompatible with our observed Markov violations. See the General Discussion for further discussion of these patterns of judgments.

One salient pattern in the data is the "spiking" at 50%. This has also been observed in between-subjects data like that from Rottman and Hastie (2016, see Figure 4A). Responses at 50% may reflect guessing or responding in some default manner. One possibility is that one of the above models, in combination with a probability of responding at 50%, can explain the observed variability. While this may account for some variance, such a model would still need to explain why the prevalence of these responses in the middle of the range differs by inference type. In particular, it has to provide an account of why those spikes are largest for inconsistent inferences and smallest for consistent inferences. One explanation could be that participants are more likely to guess when the information provided for an inference is more ambiguous.

Both response noise and uncertainty about parameters are compatible with the normative CGM framework being the underlying process used to generate causal judgments. Other models of causal reasoning predict variability as a consequence of the reasoning process itself. The mental model theory of causation stipulates that causal judgments are rendered from imagined concrete states, as determined by the causal structure that is being reasoned about (Johnson-Laird et al., 2015). A similar account from Davis & Rehder (2020) models these imagined states as being the result of a structured mental search through the space of possible situations, in the form of a Markov Chain Monte Carlo sampling process. The stochastic nature of this sampling process introduces variability. And while not explicitly designed as a process model, quantum models of causal reasoning may make unique predictions by virtue of participants varying in the dimensionality of their representations (Trueblood et al., 2017).

While all of these accounts make predictions about response distributions, the Mutation Sampler is the only model for which predictions about response distributions have been explicitly reported (Davis & Rehder, 2020). One of these predictions is that of spikes at 50% (resulting in distributions that are not unimodal), which appear to be borne out in our data. Moreover, the Mutation Sampler predicts an increase in

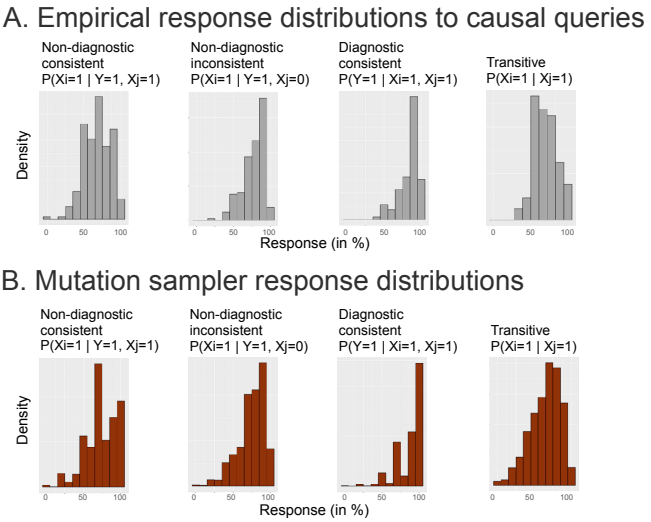

Figure 4: Comparison of A. distribution of responses from between-participant study from Rottman & Hastie (2016) and B. 10,000 sample judgments drawn from the mutation sampler (Davis & Rehder, 2020). Note: transitive inferences were not tested in this study.

spikes for inconsistent trials. The model incorporates a mechanism for default responsing at 50% when the sampling process does not provide information to answer the query. This is more likely for inferences with inconsistent information as states with incongruous variable values are sampled less often.

Discussion

This article takes the first step in bringing the field of causal reasoning in line with other domains of cognitive science that take into account the variability of judgments and not just their averages. As exemplified by the prolific use of the diffusion decision model (Ratcliff et al., 2016), response distributions provide opportunity for more sensitive signals to underlying cognitive processes. We consider the development of an experimental design that elicits repeated measurements of the same causal queries to be a major contribution of this project.

Our findings show, for the first time, that there is indeed meaningful within-participant variability in causal reasoning. That our data exhibit similar variability to that in between-subjects studies (e.g. see Figure 4A), suggests that it largely arises from the processes by which individuals generate causal inferences. That it varies with the type of causal inference supports the additional conclusion that the variability at least partly reflects a decision-making process rather than noise (e.g., noise in motor responses) or some other factor about individual participants (Rottman & Hastie, 2016).

We found that the direction of reasoning matters: Diagnostic (from effect to cause) inferences were more variable than forward (cause to effect) inferences. This squares nicely with the often repeated claim is that it is easier to think in the direction from cause to effect (Tversky & Kahneman, 1982). Our results add to the existing empirical literature on differences between diagnostic and predictive reasoning, which has reported that people take longer to respond to diagnostic queries (Fernbach & Darlow, 2010) and that they do not neglect possible alternative causes (which they tend to do for predictive inferences) (Fernbach et al., 2010). It has also been argued that diagnostic reasoning is more comparative (Fernbach et al., 2011), and CGM theory stipulates that diagnostic reasoning requires the incorporation of additional information, namely the prior probability of the cause. The fact that the more variable judgments are the ones that have been found to be more difficult suggests that the response distributions observed in this experiment are reflective of the process by which these judgments are rendered.

The information provided to participants in conditional inferences also matters: knowledge of all non-queried variables leads to an increase in variability, while incomplete information seems to reduce it. These findings are somewhat surprising. One might expect that additional information would result in less uncertainty over the possible values of an unknown variable. We find the opposite. It might be that more pieces of information result in more variability by virtue of there being more ways to process two pieces of information versus one. A related explanation appeals to stimulus encoding. When more pieces of information are provided as part of the stimulus, it might be more probable that there is more variation in whether one or more pieces of the stimulus are encoded incorrectly on a portion of the trials.

Also contrary to our expectations was the finding that, for the diagnostic judgments, inconsistent inferences were less variable than consistent ones. Indeed, Figure 1 suggests that the lower variability of inconsistent diagnostic inferences was due to the sharp clustering of responses at 50%. Note that for these inferences 50% is not only the normative answer, it may be have been especially natural for participants to use the exact middle of the response scale given that the two pieces of given information (i.e. $X_i = 1, X_j = 0$) canceled each other out (and that the base rate of the cause Y was defined to be 50%). And, as mentioned above 50% responses can also arise as a consequence of rarely sampled states (Davis & Rehder, 2020).

We also found a relationship between violations of the causal Markov condition and variability over participants. Participants who are more variable tended to exhibit stronger Markov violations. This finding squares with a large literature suggesting that Markov violations are a key source of evidence for the claim that the normative CGM framework is not an accurate model of the true underlying process that people use to draw causal judgments Rehder (2014); Rottman & Hastie (2016); Trueblood et al. (2017). Importantly, Markov violations are by definition incompatible with any model that uses the CGM framework as its core representation, and therefore defies simple interpretations of the ob-

served variability as response noise or uncertainty about the parameters of the causal model. Instead, it appears to signal that a common underlying process drives both Markov violations and part of the observed variability. This underlying process may be related to individual factors. One such factor might be a difference in reasoning strategy or style, which would be in line with findings relating Markov violations to tendency to engage less in reflective thought (Trueblood et al., 2017). Another possible factor may be limitations in working memory capacity, as proposed by Davis & Rehder (2020).

The experimental design used in this study is not without its limitations. A major experimental obstacle was eliciting 24 unique judgments for identical causal queries. Variability in judgments may have resulted from variability in interpretation of experimental materials, rather than in the causal reasoning process itself. For example, people may have different beliefs about the causal relationships between societal factors than between features of stars. We believe this possibility cannot account for all variability observed in our study (see discussion of uncertainty in parameters as a source of variability). Another limitation is that the analyses presented here use standard deviation as an index of variability. Since the distributions are not unimodal this measure does not necessarily capture all relevant information in the response distributions. Lastly, our experiment only tested a subset of the possible inferences in one particular inference task, and the extent to which our findings apply to other inferences or causal reasoning tasks is an open question.

To explain the observed variability, we discussed the correspondence between our findings and the qualitative patterns of variability in potential models of the causal reasoning process. Fitting full response distributions is a challenging computational and statistical problem that goes beyond the scope of this paper. We do wish to emphasize that future efforts should focus on this challenge, as modeling more than just averaged judgments will help improve our understanding of the cognitive processes underlying causal reasoning.

Conclusion

Causal reasoning is a core cognitive activity. Understanding the processes by which people generate causal judgments will help us better understand a range of cognitive activities from decision-making to categorization. In this paper we presented the first investigation of within-participant variability in causal judgments. This variability differs by inference type, is related to systematic reasoning errors, and is not easily explained by simple additions to the dominant CGM framework for causal inference. We hope that the data presented in this paper will broaden the scope of behavioral signals used to study how people draw causal inferences, and provide an important benchmark for formal models of this process.

References

Dasgupta, I., & Gershman, S. J. (2021). Memory as a computational resource. *Trends in Cognitive Sciences*.

Davis, Z., & Rehder, B. (2017). The causal sampler: A sampling approach to causal representation, reasoning, and learning. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (Vol. 39).

Davis, Z., & Rehder, B. (2020). A process model of causal reasoning. *Cognitive Science*, *44*.

Fernbach, P. M., & Darlow, A. (2010). Causal conditional reasoning and conditional likelihood. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (Vol. 32).

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, *21*(3), 329–336.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, *140*(2), 168.

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in cognitive sciences*, *19*(4), 201–214.

Pearl, J. (2000). *Causality: models, reasoning and inference* (Vol. 29). Springer.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in cognitive sciences*, *20*(4), 260–281.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive psychology*, *72*, 54–107.

Rehder, B. (2018). Beyond markov: Accounting for independence violations in causal reasoning. *Cognitive psychology*, *103*, 42–84.

Rehder, B., & Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & cognition*, *45*(2), 245–260.

Rottman, B. M., & Hastie, R. (2016). Do people reason rationally about causally related events? markov violations, weak inferences, and failures of explaining away. *Cognitive Psychology*, *87*, 88–134.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Trueblood, J. S., Yearsley, J. M., & Pothos, E. M. (2017). A quantum probability framework for human probabilistic inference. *Journal of Experimental Psychology: General*, *146*(9), 1307.

Tversky, A., & Kahneman, D. (1982). Causal schemas in judgments under uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (p. 117-–128). Cambridge University Press. doi: 10.1017/CBO9780511809477.009