The Mutation Sampler: A Sampling Approach to Causal Representation

Zachary J. Davis (zach.davis@nyu.edu)
Bob Rehder (bob.rehder@nyu.edu)
Department of Psychology, New York University
6 Washington Place, New York, NY, USA 10003 USA

Abstract

The causal graphical model framework has achieved success accounting for causal-based judgments across a wide variety of tasks, from reasoning to learning to interventions. However, relatively little work has investigated the processes by which people are able to make such sophisticated judgments. We propose a new process model—the *mutation sampler*—that can achieve human-level causal judgments in a psychologically plausible manner by assuming that people construct their causal representations using the Metropolis-Hastings sampling algorithm constrained to only a small number of samples (e.g., < 20). Not only does this model achieve better fits to behavior than the normative causal graphical models framework, but it can also account for systematic errors (Markov violations) that the normative model by definition cannot. The prediction that Markov violations will manifest themselves across tasks was corroborated by a new experiment that directly measured people's causal representations.

Keywords: sampling, causal representation, causal reasoning

Introduction

The representation and use of causal knowledge is a central object of investigation in the cognitive sciences. Causal models have been found to affect cognition in a wide variety of inference problems, from reasoning and learning to decision-making and categorization (for summaries, see Rottman & Hastie, 2014; Waldmann & Hagmayer, 2013). One formal model of the representation of causal information — causal graphical models — has achieved success in modeling behavior across these tasks.

While causal graphical models have been very successful at the computational level of analysis, so far there has been little work investigating how people are able to approximate the normative standard in tasks in causal learning and reasoning. Building on recent work in cognitive science that investigates the role of sampling methods in accounting for judgments in a variety of domains (Hertwig & Pleskac, 2010; Lieder, Griffiths, & Goodman, 2012; Vul et al., 2014), we propose a model for resource-constrained inference using causal models. In particular, we propose that, when reasoning about causal systems, people attend to concrete cases and shift attention between those cases systematically. This process yields a joint distribution as a representation of the causal system, which can be used for inference in any task that can be modeled with causal graphical models.

The current model fits in with the burgeoning field of resource-rational models of cognition, which explain failures to adhere to the normative model as resulting from resource limitations. In particular, sampling models have been successful in modeling systematic biases across a wide variety of tasks. Disgupta, Schulz, and Gershman (2016) showed that a wide range of effects in probability estimation can be explained by a Markov Chain Monte Carlo (MCMC) process initialized at a particular point and taking limited samples. For example, asking the probability that someone died of any disease will yield higher ratings when given examples of likely diseases (e.g. heart disease, obesity) than when given examples of unlikely diseases (e.g. muscular dystrophy, mad cow disease). This effect fits nicely with the MCMC account of anchoring proposed by Lieder, Griffiths, and Goodman (2012). In a different vein, Johnson and Busemeyer (2016) modeled diversions from expected utility theory as a limited samples Markov process with biased starting points. As in these accounts, we model causal judgments as resulting from limited samples with a biased initialization.

Just as in many other sampling accounts of cognition, our approach has two goals. Firstly, we aim to explain how people succeed at making sophisticated causal judgments. We do this by showing that a psychologically plausible number of samples can accomplish human-level causal inference. Secondly, we aim to model systematic errors, analogously to sampling accounts of the anchoring effect or prospect theory (Lieder, Griffiths, & Goodman, 2012; Johnson & Busemeyer, 2016). In causal cognition, people systematically violate the *Markov condition*, a foundational feature of causal graphical models which stipulates that the value of a node is independent of its non-descendants, conditional on its parents (Rehder, 2014). This principle is crucial for statistical inference from causal graphical models (Pearl, 1988; Koller & Friedman, 2009), and has been argued to be necessary for a rigorous account of interventions (Hausman & Woodward, 1999).

Process Model

Formalization

The proposed model is a variant of Metropolis-Hastings (MH) Markov Chain Monte Carlo, a computationally efficient rejection sampling method (Hastings, 1970). MH is defined by two components: a proposal distribution $\mathbb{Q}(q'|q)$ and a transition probability $a(q'|q)$, where q is the current state and q' is the proposal state in the random walk. Whereas MH models often deal with a continuous state space, the proposed model samples over the discrete states of a causal model. Figure 1B presents the eight states for the three variable graph shown in Figure 1A.

The sampling process uses the standard MH transition probability:

$$a(q'|q) = min\Big(1, \frac{\pi(q')}{\pi(q)}\Big)$$

where $\pi(q)$ is the joint probability of the graph being in state q given the graph's parameters (see Appendix A for an example of how $\pi(q)$ is calculated). The parameters reflect the particular beliefs of the participant (e.g. the causal strength between cause and effect).

We assume a proposal distribution $\mathbb{Q}(q'|q)$ that restricts reachable states q' to those that differ from the current state q by one binary variable. Each reachable state has an equal probability of being selected. Edges in Figure 1B denote reachable states for a node. This proposal distribution was inspired by models that assume the proposal distribution makes small adjustments to the currently held state (Bramley, Dayan, Griffiths, & Lagnado, 2017; Johnson & Busemeyer, 2016; Lieder, Griffiths, & Goodman, 2012). For direct empirical evidence for this proposal distribution, see Appendix C.

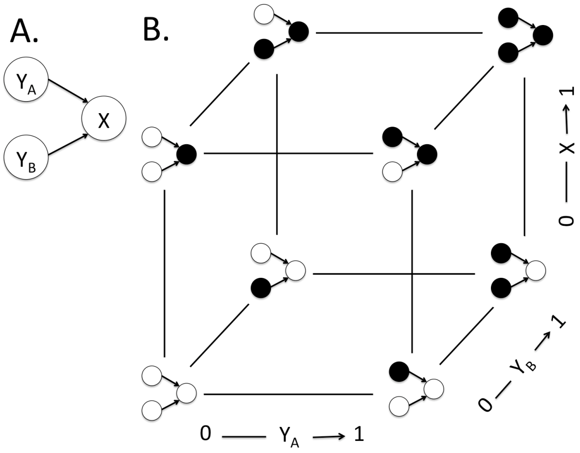


Figure 1: (A) Common effect network. (B) Possible concrete states of a common effect network. Filled in circles indicate a value of 1, empty circles indicate a value of 0.

Note that this proposal distribution confers additional efficiency benefits. Because only one variable is changed, the ratio $\frac{\pi(q')}{\pi(q)}$ simplifies to

$$\frac{\pi(v_i', v_{-i})}{\pi(v_i, v_{-i})} = \frac{\pi(v_i'|v_{-i})\pi(v_{-i})}{\pi(v_i|v_{-i})\pi(v_{-i})} = \frac{\pi(v_i'|v_{-i})}{\pi(v_i|v_{-i})}$$

where v_i is the value of node i in q, and v_i' is the value in q'. This reduces the problem to calculating the relative conditional probabilities of two states, rather than representing the entire joint distribution. That calculating conditional probabilities only requires consideration of the node's *Markov blanket* further aids efficiency (Koller & Friedman, 2009).

The model thus far is simply an efficient MH model for estimating a causal graph's joint distribution. Importantly, however, we introduce a bias in the starting point for sampling: It always starts sampling from 'prototype' states, those in which nodes are either all 0 or all 1 (bottom left and top right corners of Figure 1B). This assumption is inspired by Johnson-Laird's influential Mental Models theory, in which the most easily represented state is the one where antecedent and consequent are both true (Johnson-Laird & Byrne, 2002). We propose that prototype states are the most easily represented states of a causal graph.

state (q)	proposal (q')	$\frac{\pi(q')}{\pi(q)}$	$Unif(0,1)$	ratio > rand?
111	101	.79	.32	TRUE
101	100	.54	.74	FALSE
101	111	1.27	.84	TRUE
111	110	.21	.56	FALSE
111	101	.79	.38	TRUE
101	001	.46	.11	TRUE
001	000	2.33	.29	TRUE
000	100	.50	.33	TRUE
100	000	2.00	.80	TRUE
000	010	.50	.09	TRUE
010	...	...	...	...

Table 1: Example run of the mutation sampler for a common effect graph (causal strength = .5, causal prevalence = .5, strength of background causes = .33).

Example Run

The previous section presented the full formalization of how the mutation sampler builds a joint distribution. However, it may be helpful to see the process at work. Table 1 shows an actual run of the mutation sampler for a common effect graph. The sample run shows every step of the process. The process starts in a prototype state (either 000 or 111). Then, a proposal that differs by only one state is generated. The Hastings ratio (ratio of the proposal probability to the state probability) is calculated and compared to a random number generated from $Unif(0,1)$. If the Hastings ratio is greater than the random number, the state is updated to the proposal state. If it is smaller, the proposal is rejected. The joint distribution is built by simply summing the number of visits to each state, and dividing by the total number of states to normalize[1].

Formalization Discussion

Regardless of our proposal distribution and biased initialization, with many samples (e.g., 10^6), the mutation sampler will converge to the normative distribution. However, we assume that people are resource-constrained and thus can only take a few samples (on the order of less than twenty). Following Bramley et al. (2017), we model people as having a fixed capacity for sampling, but may vary in the number of samples taken for any particular judgment. To implement this assumption, for each judgment that a participant makes they will take k samples, where k is drawn from a Poisson distribution with mean $\lambda \in [1, \infty]$. Larger λ values signify that a participant has a capacity to take many samples, and would thus predict behavior more in line with the normative model. Smaller λ values signify a limited capacity to take samples, predicting

[1] To avoid division by zero in the case where a state was never visited, we initialize every state with a very small value (1e-10).

a stronger divergence from the normative model. With a limited capacity for taking samples, an MH model will overestimate the probability of states near the starting point (as it did not have time to fully explore the state space) and underestimate the remaining states. This effect is shown in Figure 2.

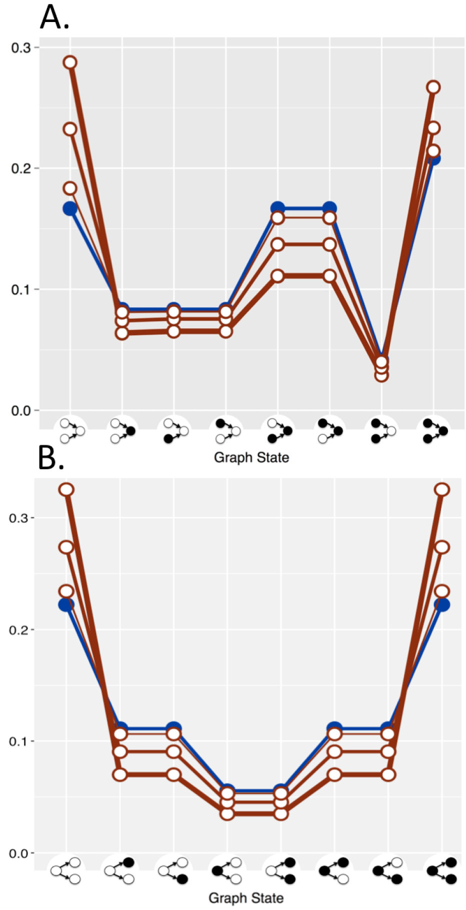


Figure 2: Simulated joint distributions for (A) Common effect (B) Common cause. Both networks parameterized with causal strength = .5, causal prevalence = .5, strength of background causes = .33. The blue lines (solid points) represents the joint distribution entailed by the normative model. Red lines (open points) represent the joint distributions simulated by the mutation sampler, with thicker lines meaning fewer samples (thick = 4 samples, medium = 8, thin = 32).

The predictions for Figure 2 represent expected values, rather than an actual run. Rather than explicit sampling, the mutation sampler's predictions for a given chain length has an analytic solution involving repeated multiplication of the matrix of transition probabilities between graph states defined by the Metropolis-Hastings rule. To generate estimated joint distributions for a capacity for taking samples (rather than a single chain length), we find the value at which the cumulative distribution of $Poiss(\lambda)$ crosses .99. Then, we calculate the expected joint distributions for each chain length up to that maximum value and weight them by the renormalized truncated Poisson to generate an expected joint distribution for a given capacity for sampling. We use these expected values when fitting participant performance.

We now test the fit of the mutation sampler to human judgments. First, we compare fits of the normative model and mutation sampler on existing datasets in causal reasoning. An important prediction of the mutation sampler model is that Markov violations are *task-general*, and thus we would not expect them to only show up in causal reasoning tasks. Instead, these violations are built into the representations of causal graphs and so will propagate to any task that uses the representation. To test this prediction, we compared our model to standard Bayes nets on a new task introduced in the "Expected Distributions" section.

Causal Reasoning

In a first empirical test of the mutation sampler, we assess how it accounts for conditional probability judgments that are drawn on the basis of causal knowledge. Appendix B presents fits of the normative model and the mutation sampler to 18 experimental conditions from four past articles, involving a total of 672 subjects and 8 distinct causal network topologies. In every condition subjects were first instructed on causal knowledge and then presented with a series of conditional probability queries. Subjects judged the presence of the to-be-inferred variable on a 0-100 scale. For each subject, we fit versions of the normative model and the mutation sampler suitable for the causal network that that subject was taught. For both models, those parameters included one representing the marginal probability of a causal node, one representing the strength of the every causal relation in the network, and one representing the strength of alternative causes (causes not explicitly part of the causal network itself), and a scaling parameter that scaled the predicted conditional probabilities onto the 0-100 rating scale. The mutation sampler had an additional fifth parameter representing the chain length, that is, the number of samples. Details of the fitting procedure and the best fitting average parameter values are presented in Appendix B for each condition along with a number of measures of quality of fit.

Appendix B reveals that the mutation sampler yielded a better fit (according to a measure that corrects for the number of parameters, AIC) as compared to the normative model in all of the 18 experimental conditions. In addition, a larger number of subjects were better fit by the mutation sampler in 15 of the 18 conditions.

We briefly summarize the performance of the mutation sampler in two key conditions. First, for the three-variable common cause condition ($Y_A \leftarrow X \rightarrow Y_B$; Figure 3A), the normative model stipulates that the two Ys should be independent conditioned on X, that is, that $p(y_i^1|x^1y_j^0) = p(y_i^1|x^1) = p(y_i^1|x^1y_j^1)$ and $p(y_i^1|x^0y_j^0) = p(y_i^1|x^0) = p(y_i^1|x^0y_j^1)$. These predictions are represented by the two horizontal blues lines in Figure 3A, which depicts the normative model's best fit to these data. The figure reveals that subjects instead judged that $p(y_i^1|x^1y_j^0) < p(y_i^1|x^1) < p(y_i^1|x^1y_j^1)$ and $p(y_i^1|x^0y_j^0) < p(y_i^1|x^0) < p(y_i^1|x^0y_j^1)$. Violations of independence with common cause networks have been found in multiple studies (Ali et al., 2011; Lagnado & Sloman, 2004; Fernbach & Rehder, 2013; Mayrhofer & Waldmann, 2015; Park & Sloman, 2013, 2014; Rehder & Burnett, 2005; Rehder, 2014a; Rottman & Hastie, 2016; Walsh & Sloman, 2004; see Hagmayer, 2016, and Rottman & Hastie, 2014, for reviews). Figure 3A also reveals that those independence violations are reproduced by the mutation sampler.

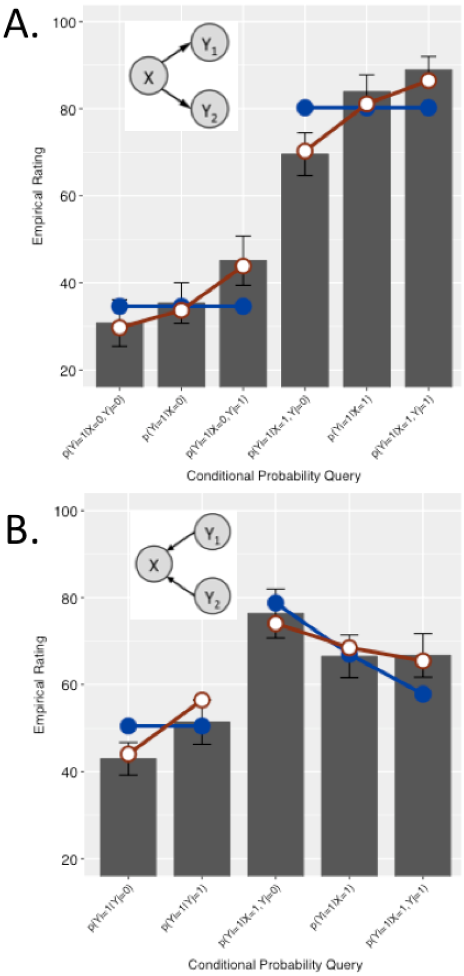


Figure 3: Data from Rehder & Waldmann (2014), Experiment 1. Sampler (red lines) and normative (blue lines, solid points) fits to conditional probability judgments. Error bars denote 95% confidence intervals.

Second, for the three-variable common effect condition ($Y_A \rightarrow X \leftarrow Y_B$; Figure 3B), the normative model stipulates that the two Ys should be unconditionally independent, that is, $p(y_i^1|y_j^0) = p(y_i^1|y_j^1)$, as depicted by the horizontal blue line in Figure 3B. Yet subjects judged that $p(y_i^1|y_j^0) < p(y_i^1|y_j^1)$ instead. This apparent expectation that the causes of a common effect graph are positively correlated has been observed in other studies (Luhmann & Ahn, 2007; Perales et al., 2004; Rehder & Burnett, 2005; Rehder, 2014a; 2014b; Rottman & Hastie, 2016; cf. Von Sydow et al., 2010). Figure 3B also reveals that the mutation sampler correctly predicts that $p(y_i^1|y_j^0) < p(y_i^1|y_j^1)$.

The mutation sampler also accounts for another reasoning error that subjects commit with the graph in Figure 3B. Explaining away is a signature property of common effect graphs. If X is observed to occur then the probability that Y_A is present of course increases. But if it is then further observed that the other cause Y_B is present then the probability that Y_A is present should decrease. (Conversely, if Y_B is observed to be absent then the probability of Y_A should increase.) In fact however, research finds that subjects often explain away too little or not at all (Morris & Larrick, 1995; Rehder, 2014; see Rottman & Hastie, 2014, for a review). The right three bars in Figure 3B illustrate the three conditional probability judgments relevant to explaining away: $p(y_i^1|x^1y_j^0), p(y_i^1|x^1)$, and $p(y_i^1|x^1y_j^1)$. The fits of the normative model to these data points reveal that explaining away with Rehder and Waldmann's subjects was indeed too weak (the slope of the blue line is steeper than the empirical ratings). In contrast, the mutation sampler correctly predicts this too weak explaining away (the slope of the red line is shallower).

Expected Distributions

Recall that when the mutation sampler's number of samples is limited, it warps a causal graph's joint distribution, overestimating prototype states and underestimating others (Figure 2). The following experiment tests this account using a novel methodology, one that directly asks participants to generate a distribution for a causal graph.

Method

Materials. Participants were presented with causal hypotheses in one of three domains: meteorology, sociology, or economics. Each domain had three variables (in economics: interest rates, trade deficits, and retirement savings; in meteorology: ozone levels, air pressure, and humidity; in sociology: urbanization, interest in religion, and socioeconomic mobility). Each variable could take on two possible values. One of these values was described as "Normal" and the other was either "High" or "Low". The values of the variables were mixed to prevent domain-specific beliefs from affecting the results (alternate values were either all "High", all "Low", or a mixture of "High" and "Low"). All hypotheses were of the form shown in Figure 1A, with two causes of one effect.

Procedure. Participants first studied screens of information that defined the variables, provided a mechanism describing how each cause could independently generate the effect, and a diagram of the causal relationships. They were then required to pass a multiple-choice test of this knowledge.

Next, participants were asked to generate a data set that they would expect to result from the causal graph. The causal relationship between smoking and lung cancer was used as an example. Subjects were shown the four cells formed by crossing smoker/non-smoker with lung cancer/no-lung cancer and how (in terms of how hypothetical people were allocated to the four cells) a greater proportion of smokers had lung cancer as compared to non-smokers. Subjects were asked to generate an analogous distribution in their assigned domain (economics, etc.). Specifically, they were given 50 pennies and asked to distribute them among the cells formed by crossing the three binary variables. They did so by placing the coins on a large sheet that contained the eight possible states (the position of the states on the sheet was randomized).

Design and Participants. The experiment consisted of a 3 (domain) by 4 (variable senses, e.g., all "High") by 2 (network structure, i.e., common cause or common effect) between-subjects design. 120 New York University undergraduates received course credit for participation.

Results

Initial analyses revealed no effect for domain or variable senses, so results were collapsed over these factors. As would naturally be expected, the allocation of coins differed depending on whether participants were generating expected distributions for common cause or common effect networks. For this reason, we will present analyses for these two groups separately.

Common effect. Figure 4 presents how subjects allocated the 50 pennies to the eight states of the graph in Figure 1A (gray bars). Because these raw data are difficult to interpret, we computed measures that reflect the statistical relationships among the three binary variables implied by the pennies. In particular, we first normalized a subject's distribution and then computed the phi coefficients between a Y and an X ($\phi(Y_i,X)$; the pennies were aggregated so that the two Ys are interchangeable), between the Ys themselves ($\phi(Y_A,Y_B)$), and between the Ys conditioned on the presence of X ($\phi(Y_A,Y_B|X)$). These measures averaged over subjects are presented in Figure 5. First, the fact that $\phi(Y_i,X) >> 0$ indicates that subjects understood that the Ys were generative causes of X. Of greater theoretical importance is the fact that $\phi(Y_A,Y_B)$ was also significantly greater than 0, $t(59) = 3.62$, $p < .001$. This corroborates our claim that the violations of independence that obtain during causal reasoning (Figure 3) are also manifested in peoples' causal representations (Figure 4).

The best fit of the normative model is shown superimposed on the empirical data in Figure 4 (blue line)[2]. The figure indicates that the normative model tends to underpredict subjects' judgments for the two prototype states (111 and 000) and overpredict the remaining states. Phi coefficients computed for these fits (blue line in Figure 5) show the expected result that the normative model requires that $\phi(Y_A,Y_B) = 0$, at odds with subjects' distributions. Moreover, it sharply underpredicts $\phi(Y_A,Y_B|X=1)$. Because of the explaining away phenomenon described above, the normative model requires that $\phi(Y_A,Y_B|X=1)$ is negative (one cause is less likely when the other is present). Figure 5 shows that subjects' distributions implied a value of $\phi(Y_A,Y_B|X=1)$ that is less negative (i.e., explaining away was again too weak).

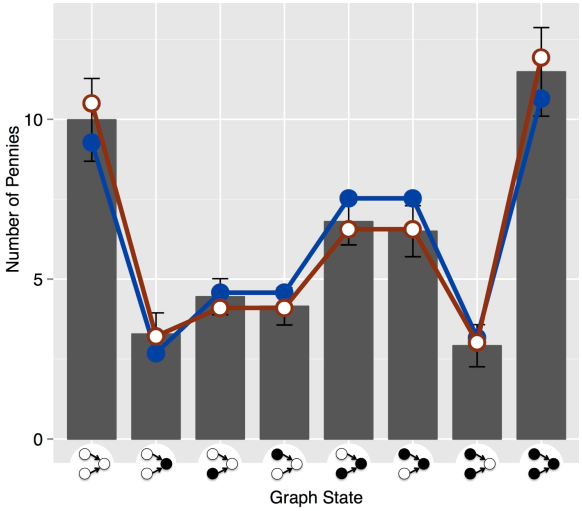


Figure 4: Common effect. Mutation sampler (red line) and normative (blue line, solid points) fits to participant-generated expected distribution judgments. Error bars denote 95% confidence intervals.

The best fit of the mutation sampler (red lines in Figs. 4 and 5) shows that it accounts for the fact that, relative to the normative model, the number of pennies is generally too large for the prototype states and too small for other states[3]. Of course, this pattern was expected on the basis of the theoretical predictions in Figure 2. Like the subjects, the mutation sampler predicts that $\phi(Y_A,Y_B) > 0$ and that explaining away (as represented by $\phi(Y_A,Y_B|X=1)$ is too weak relative to the normative model. As a result, it achieved a better fit to these data than the normative model (*AIC* of 3.6 vs. 10.8).

Common cause. Figure 6 presents how subjects allocated the 50 pennies to the eight states of a common cause graph (one cause, two effects). As for the common effect structure, we computed measures that reflect the statistical relationships among the three binary variables implied by the pennies. In particular, we first normalized a subject's distribution and then computed the phi coefficients between a Y and an X ($\phi(Y_i,X)$; the pennies were aggregated so that the two Ys are interchangeable), between the Ys themselves ($\phi(Y_A,Y_B)$), and between the Ys conditioned on the presence of X ($\phi(Y_A,Y_B|X=1)$). These measures averaged over subjects are presented in Figure 7. First, the fact that $\phi(Y_i,X) >> 0$ indicates that subjects understood that X was a generative cause of the Ys. Of greater theoretical importance is the fact that $\phi(Y_A,Y_B|X)$ was also significantly greater than 0, $t(59)$

[2] The best fitting parameters (w_Y = .519, w_{YX} = 0.440, w_X = .243 averaged over subjects), were those that maximized the likelihood of the distribution of pennies.

[3] $w_Y = .534, w_{YX} = 0.410, w_X = .328$, chain length = 10.1.

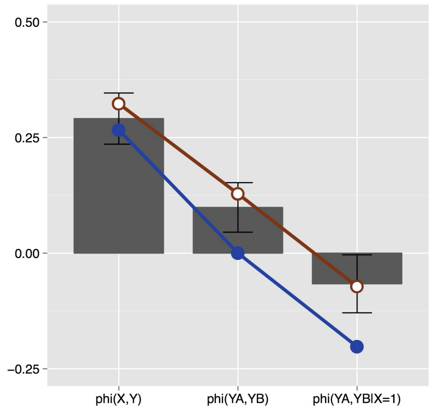


Figure 5: Common effect. Mutation sampler (red line) and normative (blue line, solid points) fits to participant-generated expected distribution judgments. Error bars denote 95% confidence intervals.

= 3.60, $p < .001$. This corroborates our claim that the violations of independence that obtain during causal reasoning (Figure 3) are also manifested in peoples' causal representations (Figure 6).

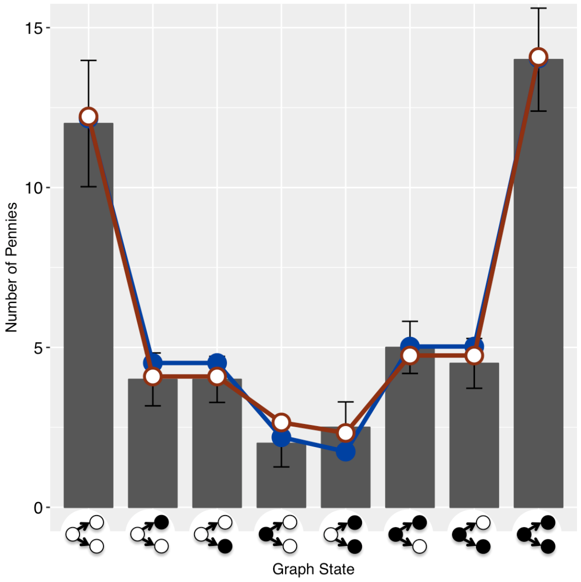


Figure 6: Common cause. Mutation sampler (red line) and normative (blue line, solid points) fits to participant-generated expected distribution judgments. Error bars denote 95% confidence intervals.

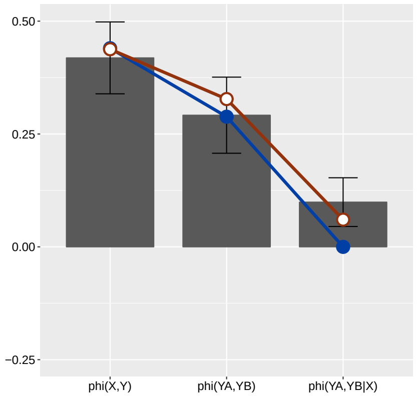


Figure 7: Common cause. Mutation sampler (red line) and normative (blue line, solid points) fits to participant-generated expected distribution judgments. Error bars denote 95% confidence intervals.

Figure 6 also shows fits for the normative and mutation sampler models to participant data (blue and red lines, respectively). While the fits of two models are visually much closer for the common cause network than they were for the common effect network, the mutation sampler still fit participant judgments more closely than the normative model (*AIC* of -14.08 vs. -23.03)[4]. Qualitatively, Figure 7 shows that the normative model cannot account for the Markov violations exhibited by a $\phi(Y_i, Y_j|X)$ significantly greater than 0, whereas the mutation sampler has no such restrictions.

Discussion

Although causal graphical models have enjoyed success in explaining causal cognition, relatively little work has investigated the cognitive processes by which such sophisticated judgments are made. We introduced a new process model that yields a better fit across a diverse array of experimental conditions encompassing 672 subjects and confirmed a key prediction of the model on another 120 subjects using a new methodology that assessed, in a relatively direct way, people's causal representations. The results suggest that the fault lies not in how we reason or learn but how we represent.

This paper has proposed a process model that naturally constructs faulty causal representations. Importantly, it does so in a manner that is computationally efficient and psychologically plausible. The Metropolis-Hasting rule combined with the proposal distribution we advocate implies that at any one time reasoners only need to consider the relative likelihood of two graph states that differ by one variable, a computation that can be carried out very efficiently (because it involves only those nodes in the variable's Markov blanket; Koller & Friendman, 2009). Yet further efficiencies can be achieved for conditional probability queries (because sampling can be limited to those graph states that instantiate a query's antecedent). Note that this view suggests that humans *could* construct veridical causal representations—if only they had the cognitive resources to do so. The fault thus lies not in our causal representations per se but rather in the fact

[4] Normative best fitting parameters: $w_X = .510, w_{XY} = .556, w_Y = .285$. Mutation sampler best fitting parameters: $w_X = .472, w_{XY} = .466, w_Y = .323$, chain length = 8.3.

that causal judgments must be computed in finite time and with limited resources. Independence violations are thus an unavoidable consequence of the tradeoff between accuracy, speed, and effort.

The mutation sampler perhaps gains some credence given the property it shares with the well-known Mental Model theory, namely, that reasoning is based on concrete states of the world (Goldvarg & Johnson-Laird, 2001; Johnson-Laird & Byrne, 2002). There are, however, some differences. Whereas the model theory never represents cause-present/effect-absent situations, the mutation sampler, as a probabilistic model, merely asserts that such situations are unlikely (depending on the causal graph's parameters) and thus rarely sampled (cf. Khemlani, Barbey, & Johnson-Laird, 2014). There are also differences regarding which states reasoners initially consider (initial mental models are similar but not identical to the mutation sampler's starting samples).

There are many possible directions for future research. For one, current models do not attempt to model the substantial variability in people's causal inferences (Rehder, 2014; Rottman & Hastie, 2016). The stochastic nature of sampling may shed light on this important aspect of behavior. The mutation sampler also makes predictions about reaction times. For example, it would predict that longer reaction times implies a less warped joint distribution (because more samples were taken).

Research in the causal graphical model tradition has rarely considered the cognitive processes involved in causal-based judgments. A limited sampling approach to building causal representations (a) is psychologically plausible, (b) accounts for the key discrepancy between graphical models and human judgments (Markov violations), and (c) explains why those discrepancies manifest themselves in multiple causal-based tasks. Yet, it doesn't deny that people are sophisticated causal reasoners—they are, however, limited ones. As a process model, the mutation sampler allows the causal graphical model framework to be extended to new phenomena, such as within- and between-subject variability and response times.

References

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, 124(3), 301.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25*(4), 565-610.

Hausman, D. M., & Woodward, J. (1999). Independence, invariance and the causal Markov condition. *The British journal for the philosophy of science*, *50*(4), 521-583.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*, 97-109.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition, 115*

Johnson, J. G., & Busemeyer, J. R. (2016). A Computational Model of the Attention Process in Risky Choice. *Decision*, advance online.

Johnson-Laird, P. N., & Byrne, R. M. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review, 109*(4), 646.

Khemlani, S. S., Barbey, A. K., & Johnson-Laird, P. N. (2014). Causal reasoning with mental models. *Frontiers in Human Neuroscience, 8*, 849.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. In P. PBartlett, et al. (Eds.), *Advances in neural information processing systems* (Vol. 25, pp. 2699-2707). Cambridge, MA: MIT Press.

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review, 102*, 331-355.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

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

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of object categories. *Cognitive Psychology, 50*, 264-314.

Rehder, B., & Davis, Z. (2016). Evaluating causal hypotheses: The curious case of correlated cues. In Papafragou, A., et al. (Eds.) (2016). *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Rehder, B., & Waldmann, M. R. (2016). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory and Cognition*, 1-16.

Rottman, B., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*.

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.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*, 599-637.

Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed.), *Oxford Handbook of Cognitive Psychology* (pp. 733-752). New York: Oxford University Press.

Appendix A

For a common cause network, we show how to calculate the joint probability of a state, $\pi(X,Y_1,Y_2)$. According to the Markov condition, $\pi(X,Y_1,Y_2)$ factors such that

$$\pi(X,Y_1,Y_2) = p(Y_1|X)p(Y_2|X)p(X)$$

Assuming that causes are generative and combine according to a noisy or integration function (Cheng, 1997),

$$\pi(y_1^1|X) = 1-(1-w_{Y_1})(1-w_{XY_1})^{ind(X)}$$

$$= w_{XY_1} + w_{Y_1} - w_{XY_1}w_{Y_1} \text{ when } X = 1$$

$$= w_{Y_1} \text{ when } X = 0$$

$$\pi(y_2^1|X) = 1-(1-w_{Y_2})(1-w_{XY_2})^{ind(X)}$$

$$= w_{XY_2} + w_{Y_2} - w_{XY_2}w_{Y_2} \text{ when } X = 1$$

$$= w_{Y_2} \text{ when } X = 0$$

where w_{XY_1} and w_{XY_2} are the strengths or causal powers (Cheng, 1997) of the $x^1 \rightarrow y_1^1$ and $x^1 \rightarrow y_2^1$ causal relationships, respectively, w_{Y_1} and w_{Y_2} are the aggregate strengths of exogenous causes (causes external to the network) of Y_1 and Y_2, respectively, and ind is an indicator function that evaluates to 1 when a variable is present and 0 otherwise. That the presence of Y_i given the presence of X, $\pi(y_i^1 x^1)$, equals $w_{XY_i} + w_{Y_i} - w_{XY_i}w_{Y_i}$ makes clear that each Y_i can be brought about by X (w_{XY_i}) or its exogenous causes (w_{Y_i}). When X is absent the probability of Y_i of course is just w_{Y_i}. w_X is defined as the strength of extrinsic causes of X, which, because X has no explicit within-network causes, corresponds to X's marginal probability (i.e., $\pi(x^1)$).

Appendix B: Model Fits to Conditional Reasoning Studies

The normative model and the mutation sampler were fit to causal reasoning data from 18 experimental conditions reported in four articles. See Figure 10 for full breakdown.

Rehder and Burnett (2005) taught subjects categories whose features were causally related and then asked subjects to judge the probability that a category member had a feature given a number of the category member's features. In every experiment categories had four features. Experiment 1 tested a common cause network (hereafter referred to as CC network) in which a single feature caused three others. Conditional probability queries presented a category member in which the state of three features was given and asked subject to predict a third feature. Experiment 3 was identical to Experiment 1 except that there were two rather than three "given" features in the conditional probability queries. Experiment 4 was identical to Experiment 3 except that the four features formed a common effect network a single feature was caused by three others. The three features were described as independent causes of the effect (thus, a CE-Indep. network). In Experiment 5 the features formed a causal chain (CH) and there were three given features in the conditional probability queries.

Rehder (2014b) also tested causally-related category features. All categories had six features. In Experiment 1, three features formed a common effect network in which two features independently caused a third (CE-Indep.). The other three features formed a common effect network in which two features conjunctively caused a third (CE-Conj.). Experiment 2 tested two between-subject conditions. In one, the two triplets of features each formed an independent common effect network but in one triplet the causal relations were described as only operating "occasionally" (CE-Indep. [weak]) whereas in the other they were described as operating "often" (CE-Indep. [strong]). In the other between subject the triplets each formed a conjunctive common effect network and the links were either weak (CE-Conj. [weak]) or strong (CE-Conj. [strong]).

Rehder and Waldmann (2016) instructed subjects on two causal relationships taken from the domain of either economics, sociology, or meteorology. Those relationships formed a common effect network in Experiment 1 and a common cause network in Experiment 2. Each experiment compared between-subject conditions in which subjects were given either just a causal network, just data generated from that causal network, or both; here we present fits to the conditions in which subjects were only given a causal network.

Using the same materials as Rehder and Waldmann (2016), Rehder (2017, submitted) conducted three experiments that each tested an extended common cause network (in which the effects themselves had effects: $Z_A \leftarrow Y_A \leftarrow X \rightarrow Y_B \rightarrow Z_B$) and an extended common effect network (in which the causes themselves had causes: $Z_A \rightarrow Y_A \rightarrow X \leftarrow Y_B \leftarrow Z_B$). The number of distinct types conditional probabiity judgments was 19 in Experiment 1 and 27 in Experiments 2 and 3. In Experiment 3 only subjects were told that the causal relations operated 75% of the time.

In each condition subjects' ratings were fit according to

$$rating(t_i) = s * p_{M,\theta_M}(t_i) \tag{1}$$

where t_i is a test items, M is a model, θ_M are its parameters. The normative model and the mutation sampler had four parameters in common. Parameter c represents the base rate of a cause node in a causal network (a variable that itself has no explicit cause in the network). Parameter m represents the strength of the causal links in the network. Parameter b represents the strength of the caual links in the network. Parameters c, m, and b represent probabilities are were constrained to (0, 1). Parameter s (constrained to the range 0-300) is a scaling parameter that maps M's predictions onto the 0-100 response scale. The mutation sampler included an additional fifth parameter representing the chain length, which was constrained to the range [1/64, 1/2]. The models were fit to each subject's causal judgments by identifying parameters that minimized squared error.

For each experimental condition, Figure 10 presents the number of variables in the network, the number of subjects tested in that condition, and the number of distinct types of conditional probability queries they answered. For both the normative model and the mutation sampler it also presents the model's best fitting parameters averaged over subjects and a number of measures of fit, including the correlation between predicted and observed values (R), and a measure (*AIC*) that takes into account a model's number of parameters[5], each averaged over the subjects in that condition. Finally, for each condition the last column of Figure 10 presents the percentage of subjects best fit by the mutation sampler.

[5] *AIC = nlog (SSE/n) + 2(p + 1)* where *SSE* = sum of squared error for a participant, n = number of data points fit (30), and p = a model's number of parameters. This measure was deemed by Burnham and Anderson (1998) as appropriate for comparing models fit by least squares.

Appendix C: Analysis of the Mutation Sampler's Proposal Distribution

The extended common cause and common effect networks in Figures 8A and 8B provide an opportunity to test the mutation sampler's proposal distribution. Recall that that proposal distribution specifies that only one variable can be mutated to generate a new proposal. This will result in a potential transition to a "neighboring state", one that only differs by a single variable. To demonstrate the importance of this proposal distribution, we defined an alternative version of the mutation sampler without this proposal distribution, that is, one in which many variables could be simultaneously mutated. Figures 8C and D present the fit of this alternative model to the data in Figures 8A and 8B and shows that while it can account for some of the independence violations ($p(z^1_i|x^0y^0_iy^0_jz^0_i) < p(z^1_i|x^0y^0_iy^0_jz^1_i)$ and $p(z^1_i|x^1y^1_iy^1_jz^0_i) < p(z^1_i|x^1y^1_iy^1_jz^1_i)$), it cannot account for others ($p(z^1_i|x^0y^0_iy^1_jz^0_i) \approx p(z^1_i|x^0y^0_iy^1_jz^1_i)$ and $p(z^1_i|x^1y^1_iy^0_jz^0_i) \approx p(z^1_i|x^1y^1_iy^0_jz^1_i)$). What characterizes the latter two pairs of conditional probabilities is that their computation requires only elements of the joint distribution that involve network states with mixed 1s and 0s. However, in the absence of the mutation sampler's proposal distribution mixed networks states do not become more or less probable relative to one another, and thus the resulting conditional probabilities remain unchanged (i.e., conditional independence obtains). What the fits in Figures 8A and 8B establish is that a mutation sampler with the proposal distribution we have specified is needed to provide a full account of the observed pattern of independence violations.

Why does the neighbors proposal succeed where the sampler without this proposal distribution fails?

For the extended common cause network in Figure 8A, Table 2 presents four key conditional probabilities derived from the normative model, the mutation sampler, and the mutation sampler without the neighbors proposal distribution. The parameters were c = .50, m=.75, and b=.25; for the two sampling models the chain length was 16. The table confirms that whereas both sampling models account for the independence violation $p(z^1_i|x^1y^1_iy^1_jz^0_j) < p(z^1_i|x^1y^1_iy^1_jz^1_j)$, only the mutation sampler with the neighbors proposal distribution accounts for the independence violation $p(z^1_i|x^1y^1_iy^0_jz^0_j) < p(z^1_i|x^1y^1_iy^0_jz^1_j)$.

To demonstrate the reason for these differences, Figure 9 presents, for each of the four conditional probabilities in Table 2, the two joint probabilities that are involved in its computation. Both panels present the joint probabilities specified by the normative model (blue lines), Panel A presents those probabilities for the mutation sampler (red lines), and Panel B presents those probabilities for the mutation sampler without the neighbors proposal distribution (green lines). Each pair of joint probabilities has been normalized in order to compare their relative magnitude (which is all that matters for the computation of the corresponding conditional probability).

Both panels show that each pair of joint probabilities predicted by the normative model has the same ratio accounting for the four equal conditional probabilities in Table 2 predicted by that model. In contrast, Panel A shows that, for the mutation sampler, the ratio between $p(x^1y^1_iy^1_jz^1_iz^1_j)andp(x^1y^1_iy^1_jz^0_iz^1_j)$ (which determines the conditional probability $p(z^1_i|x^1y^1_iy^1_jz^1_j)$) is larger than the ratio between $p(x^1y^1_iy^1_jz^1_iz^0_j)$ and $p(x^1y^1_iy^1_jz^0_iz^0_j)$ (which determines the conditional probability $p(z^1_i|x^1y^1_iy^1_jz^0_j)$). Thus, $p(z^1_i|x^1y^1_iy^1_jz^0_i) < p(z^1_i|x^1y^1_iy^1_jz^1_j)$, that is, an independence violation obtains. Analogously, it shows that $p(x^1y^1_iy^0_jz^1_iz^0_j)/p(x^1y^1_iy^0_jz^0_iz^0_j) < p(x^1y^1_iy^0_jz^1_iz^1_j)/p(x^1y^1_iy^0_jz^0_iz^1_j)$ and thus $p(z^1_i|x^1y^1_iy^0_jz^0_i) < p(z^1_i|x^1y^1_iy^0_jz^1_j)$. These patterns of ratios in Panel A explain the patterns of conditional probabilities (and independence violations) the mutation sampler predicts in Table 2.

In contrast, Panel B shows that whereas the mutation sampler without the neighbors proposal distribution predicts that $p(x^1y^1_iy^1_jz^1_iz^1_j)/p(x^1y^1_iy^1_jz^0_iz^1_j) < p(x^1y^1_iy^1_jz^1_iz^0_j)/p(x^1y^1_iy^1_jz^0_iz^0_j)$ (and thus an independence violation: $p(z^1_i|x^1y^1_iy^1_jz^0_i) < p(z^1_i|x^1y^1_iy^1_jz^1_j)$), it predicts that $p(x^1y^1_iy^0_jz^1_iz^0_j)/p(x^1y^1_iy^0_jz^0_iz^0_j) = p(x^1y^1_iy^0_jz^1_iz^1_j)/p(x^1y^1_iy^0_jz^0_iz^1_j)$ (and thus the absence of an independence violation: $p(z^1_i|x^1y^1_iy^0_jz^0_i) = p(z^1_i|x^1y^1_iy^0_jz^1_j)$). This pattern arises because whereas the bias to begin sampling at the prototypes overweighs the joint probability of the two prototypes, the neighbors proposal distribution is required to result in changes to the relative probabilities of network states that don't involve the prototypes. Without the neighbors proposal distribution, any conditional probability whose computation doesn't involve the prototypes is the same as that specified by the normative model.

Judgment	Normative Model	Mutation sampler wo/ neighbors proposal distribution	Mutation sampler w/ neighbors proposal distribution
$p(z^1_i\|x^1y^1_iy^0_jz^0_i)$	.8125	.8125	.7976
$p(z^1_i\|x^1y^1_iy^0_jz^1_i)$	.8125	.8125	.8348
$p(z^1_i\|x^1y^1_iy^1_jz^0_i)$	.8125	.8125	.8436
$p(z^1_i\|x^1y^1_iy^1_jz^1_i)$	.8125	.9106	.8825

Table 2: Four key conditional probabilities for the extended common cause network as predicted by three models.

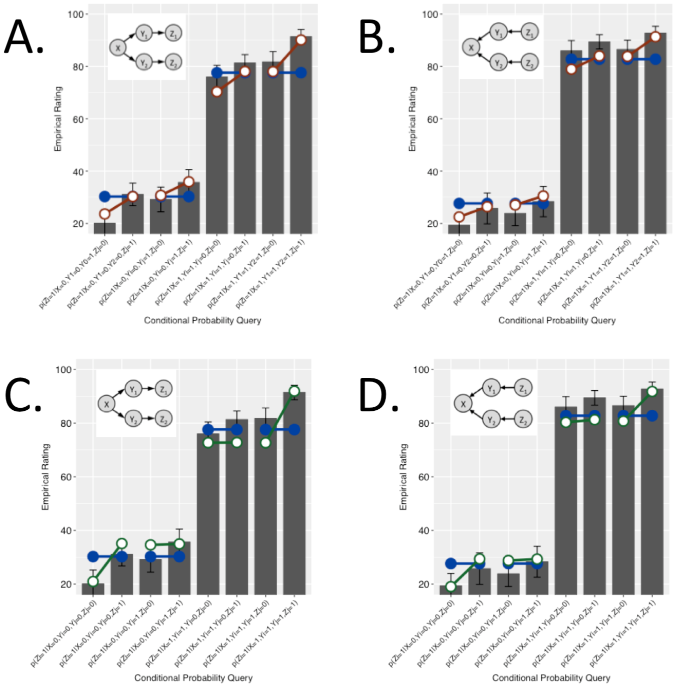


Figure 8: Empirical analysis of proposal distribution. (A) Extended common cause fits of normative model (blue lines, solid points) and mutation sampler with proposal distribution (red lines, open points). (B) Extended common effect fits of normative model (blue lines, solid points) and mutation sampler with proposal distribution (red lines, open points). (C) and (D) mirror (A) and (B), respectively, except that green lines, open points represent the mutation sampler without a proposal distribution.

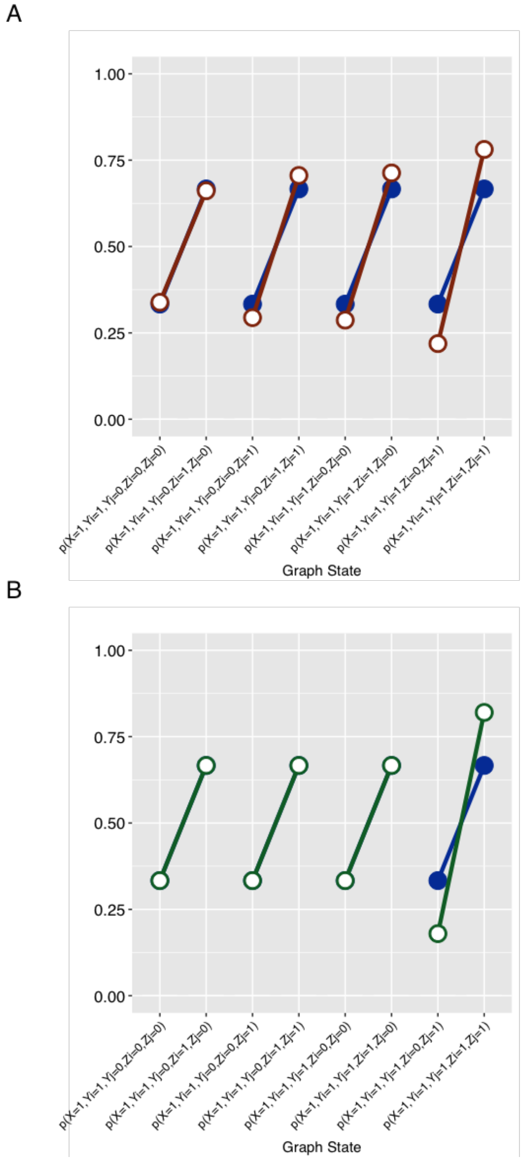
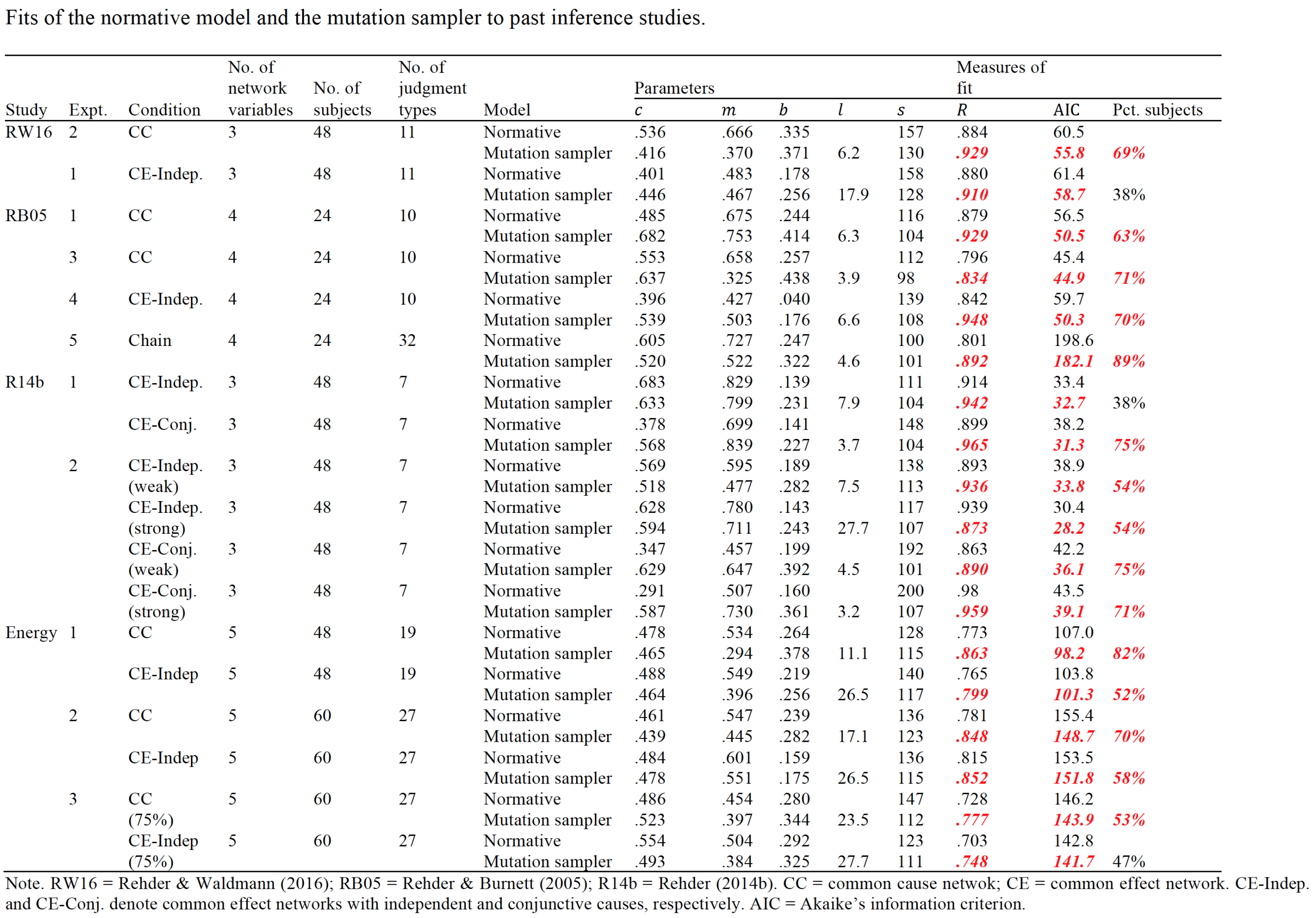


Figure 9: Graphical prediction of the implications for the three models displayed in Table 2



Fits of the normative model and the mutation sampler to past inference studies.

Study	Expt.	Condition	No. of network variables	No. of subjects	No. of judgment types	Model	Parameters					Measures of fit		
							c	*m*	*b*	*l*	*s*	*R*	AIC	Pct. subjects
RW16	2	CC	3	48	11	Normative	.536	.666	.335		157	.884	60.5	
						Mutation sampler	.416	.370	.371	6.2	130	***.929***	***55.8***	***69%***
	1	CE-Indep.	3	48	11	Normative	.401	.483	.178		158	.880	61.4	
						Mutation sampler	.446	.467	.256	17.9	128	***.910***	***58.7***	38%
RB05	1	CC	4	24	10	Normative	.485	.675	.244		116	.879	56.5	
						Mutation sampler	.682	.753	.414	6.3	104	***.929***	***50.5***	***63%***
	3	CC	4	24	10	Normative	.553	.658	.257		112	.796	45.4	
						Mutation sampler	.637	.325	.438	3.9	98	***.834***	***44.9***	***71%***
	4	CE-Indep.	4	24	10	Normative	.396	.427	.040		139	.842	59.7	
						Mutation sampler	.539	.503	.176	6.6	108	***.948***	***50.3***	***70%***
	5	Chain	4	24	32	Normative	.605	.727	.247		100	.801	198.6	
						Mutation sampler	.520	.522	.322	4.6	101	***.892***	***182.1***	***89%***
R14b	1	CE-Indep.	3	48	7	Normative	.683	.829	.139		111	.914	33.4	
						Mutation sampler	.633	.799	.231	7.9	104	***.942***	***32.7***	38%
		CE-Conj.	3	48	7	Normative	.378	.699	.141		148	.899	38.2	
						Mutation sampler	.568	.839	.227	3.7	104	***.965***	***31.3***	***75%***
	2	CE-Indep. (weak)	3	48	7	Normative	.569	.595	.189		138	.893	38.9	
						Mutation sampler	.518	.477	.282	7.5	113	***.936***	***33.8***	***54%***
		CE-Indep. (strong)	3	48	7	Normative	.628	.780	.143		117	.939	30.4	
						Mutation sampler	.594	.711	.243	27.7	107	***.873***	***28.2***	***54%***
		CE-Conj. (weak)	3	48	7	Normative	.347	.457	.199		192	.863	42.2	
						Mutation sampler	.629	.647	.392	4.5	101	***.890***	***36.1***	***75%***
		CE-Conj. (strong)	3	48	7	Normative	.291	.507	.160		200	.98	43.5	
						Mutation sampler	.587	.730	.361	3.2	107	***.959***	***39.1***	***71%***
Energy	1	CC	5	48	19	Normative	.478	.534	.264		128	.773	107.0	
						Mutation sampler	.465	.294	.378	11.1	115	***.863***	***98.2***	***82%***
		CE-Indep	5	48	19	Normative	.488	.549	.219		140	.765	103.8	
						Mutation sampler	.464	.396	.256	26.5	117	***.799***	***101.3***	***52%***
	2	CC	5	60	27	Normative	.461	.547	.239		136	.781	155.4	
						Mutation sampler	.439	.445	.282	17.1	123	***.848***	***148.7***	***70%***
		CE-Indep	5	60	27	Normative	.484	.601	.159		136	.815	153.5	
						Mutation sampler	.478	.551	.175	26.5	115	***.852***	***151.8***	***58%***
	3	CC (75%)	5	60	27	Normative	.486	.454	.280		147	.728	146.2	
						Mutation sampler	.523	.397	.344	23.5	112	***.777***	***143.9***	***53%***
		CE-Indep (75%)	5	60	27	Normative	.554	.504	.292		123	.703	142.8	
						Mutation sampler	.493	.384	.325	27.7	111	***.748***	***141.7***	47%

Note. RW16 = Rehder & Waldmann (2016); RB05 = Rehder & Burnett (2005); R14b = Rehder (2014b). CC = common cause netwok; CE = common effect network. CE-Indep. and CE-Conj. denote common effect networks with independent and conjunctive causes, respectively. AIC = Akaike's information criterion.

Figure 10: Comparison of normative model and mutation sampler for 4 existing datasets in causal reasoning.