ORIGINAL ARTICLE

Journal Section

A Process Model of Causal Cognition

Zachary J. Davis[1] | **Bob Rehder**[1]

[1]Department of Psychology, New York University, 6 Washington Place, New York, NY 10003 USA

Correspondence
Zachary J. Davis, Department of Psychology, New York University
Email: zach.davis@nyu.edu

Funding information

How do we make causal judgments? Many studies have demonstrated that people are capable causal reasoners, achieving success on tasks from reasoning to categorization to interventions. However, less is known about the mental processes people use to achieve such sophisticated judgments. We propose a new process model—the *mutation sampler*—that models causal judgments as based on a *sample* of possible states of the causal system generated using the Metropolis-Hastings sampling algorithm. Across a diverse array of tasks and conditions encompassing over 1,700 participants, we found that our model provided a consistently closer fit to participant judgments than standard causal graphical models. In particular, we found that the biases introduced by mutation sampling accounted for people's consistent, predictable errors that the normative model by definition could not. Moreover, using a novel methodology, we found that those biases appeared in the samples that participants explicitly judged to be representative of a causal system. We advocate the mutation sampler as the heretofore missing process level account of the computations specified by the causal graphical model framework, and highlight opportunities for future research to identify not just *what* reasoners compute when drawing causal inferences, but also *how* they compute it.

KEYWORDS
sampling, causal representation, causal reasoning, process models

1 | INTRODUCTION

The representation and use of causal knowledge is a central object of investigation in the cognitive sciences. Causal knowledge has 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 knowledge—causal graphical models—has achieved success in modeling human performance across these tasks. A well-known advantage of causal graphical models is that they provide a compact representation of a causal system—only the local relations between a variable and its causal parents need be explicitly represented. Another is that they are accompanied by formal methods that specify how a causal model should be learned from observed data, used to draw inferences (including counterfactual judgments and the effects of interventions by an external agent), and updated in light of changing conditions (e.g. a malfunctioning component). However, causal graphical models are understood to provide a *computational level* account of causal cognition—like all such accounts they specify *what* but not necessarily *how* specific computations are carried out (Anderson, 1990; Marr, 1982). This article extends this past work by proposing a new *rational process model* of the cognitive mechanisms that underlie many causal judgments. As a process model, the goals of this account include explaining why people commit the causal reasoning errors they do and how the correct inferences they draw can be computed within the resource limitations imposed by the human cognitive system.

1.1 | Sampling

Bayesian modeling has provided an influential account of human cognition (Griffiths et al., 2010). However, because these models are often highly computationally expensive, they are generally regarded as computational level accounts of behavior. Recently, a major project has been under way in the cognitive sciences to identify the processes by which people are able to approximate the normative Bayesian standard. Generally, researchers look for models that are minimally resource intensive but will still, at asymptote, converge to the correct answer. Monte Carlo methods are a natural approach to this resource-accuracy tradeoff. In particular, this paper deals with a popular variant of Monte Carlo methods—Markov chain Monte Carlo (MCMC)—that has achieved particular success in modeling cognition. For more extensive treatment of Monte Carlo methods and their uses in cognitive science, see Dasgupta, Schultz, and Gershman (2017).

MCMC models have successfully accounted for systematic biases across a variety of tasks. In one of the first applications of sampling methods to cognition, Lieder, Griffiths, and Goodman (2012) showed that a simple MCMC model replicates the classic anchoring and adjustment effect (Tversky & Kahneman, 1974). It has been argued that taking a limited number of samples can be rational when taking into account things such as time costs (Vul, Goodman, Griffiths, & Tenenbaum, 2014) or the utility of exaggerated differences between decisions (Hertwig & Pleskac, 2010). Dasgupta et al. (2017) proposed MCMC models as a unified account of how people approximate Bayes' rule, capturing diverse phenomena such as the crowd within (Vul & Pashler, 2008) or self-generation (Koehler, 1994) effects. Sampling accounts are not restricted to probability estimation tasks. For example, Johnson and Busemeyer (2016) modeled deviations from expected utility theory as resulting from biased sampling from prospects.

1.2 | Sampling and Causal Models

We propose a model for resource-constrained inference using causal graphical models. In particular, we propose that, when reasoning about a causal system, people think about concrete cases—states of the causal system in which all

relevant variables are instantiated with values. For example, consider the causal graph in Figure 1A in which variables Y_A and Y_B are causes of variable X. Because the variables in this network are assumed to be binary, the *state space* of this graph—the possible assignments to the three variables—consists of the eight states shown in Figure 1B. Our model assumes that reasoners sequentially *sample* these states. Later we will show that these generated samples can be used to carry out the kinds of inferences 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. As in other sampling accounts of cognition, our approach is a balancing act between two goals. On one hand, 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 reproduce human-level causal inference. On the other hand, we also aim to model the consistent, predictable errors commonly observed in research on causal cognition, analogously to sampling accounts of the anchoring effect or prospect theory (Lieder, Griffiths, & Goodman, 2012; Johnson & Busemeyer, 2016). For example, people systematically violate the *Markov condition*, a foundational feature of causal graphical models that defines patterns of conditional independence among a graph's variables. 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). We describe the Markov condition and the empirical violations later.

The paper will proceed as follows. We first formalize our model, which we dub the *mutation sampler*, and give justification for its commitments. We then compare fits of the mutation sampler and standard causal graphical models to empirical data from a wide variety of tasks and conditions. In particular, we compare fits to existing studies of causal reasoning, categorization, and intervention decision-making. The final section of the paper introduces a new experimental paradigm that assesses whether the properties of samples generated by the mutation sampler can also be observed in the samples of causal systems that reasoners generate themselves. In the General Discussion we will compare the mutation sampler to other models of causal-based judgments.

To foreshadow, we show that the mutation sampler consistently provides a better fit to multiple types of causal judgments as compared to normative causal graphical models. This finding demonstrates that both the striking successes and systematic errors that people make when reasoning about causal systems can be accounted for by the assumption that they reason on the basis of a modest number of samples taken from the possible states of a causal system rather than computing a normative response. We will conclude by advocating the mutation sampler as the heretofore missing process level account of the computations specified by the causal graphical model framework.

2 | THE MUTATION SAMPLER

2.1 | Formalization

The proposed model is a variant of Metropolis-Hastings (MH) Markov Chain Monte Carlo, a computationally efficient rejection sampling method for estimating probability distributions (Hastings, 1970; van Ravenzwaaij, Cassey, & Brown, 2018). MH methods sample from a distribution in a manner that ensures that the generated samples will, after normalization, approximate the original distribution, with convergence guaranteed as the size of the sample grows large. Whereas MH models often deal with a continuous state space, the proposed model samples over the discrete states of a causal graph — like the one in Figure 1. Just as with any distribution, the joint probability distribution associated with this graph can be approximated via MH sampling over its states. Note that our psychological claim will be that people make causal judgments on the basis of the samples they draw from a causal graph without necessarily forming a normalized joint distribution. The unnormalized joint distribution represented by the sample is sufficient to model the

inferences that can be computed from a causal graphical model.

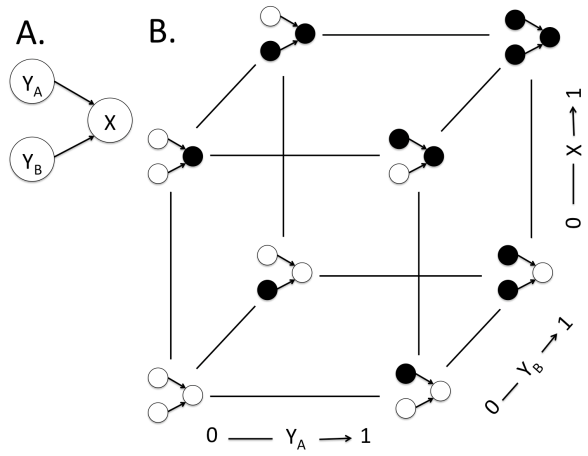


FIGURE 1 (A) A common effect graph. (B) Possible states of a common effect graph. Filled circles indicate a variable instantiated with a value of 1, open circles one with a value of 0. Edges denote reachable states as defined by the mutation sampler's proposal distribution.

Like all MCMC methods, MH constructs a sequence (or chain) of samples, where the choice of each subsequent sample in the chain depends on the previous sample. Specifically, 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 a random walk. The standard MH transition probability $a(q'|q)$ determines whether the next state in the chain should repeat the current state q or involve a transition to the new state q' and is defined by,

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

where $\pi(q)$ is the joint probability of the causal system being in state q. Importantly, $a(q'|q)$ only requires the computation of the *relative* probability of two system states, $\pi(q')$ and $\pi(q)$. This property is key because it means that $a(q'|q)$ can be computed without access to the graph's full joint distribution, which of course is what the sampling process is attempting to approximate.

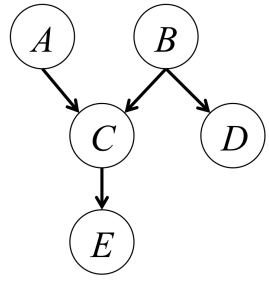
The proposal distribution $\mathbb{Q}(q'|q)$ determines which graph state should be proposed as the next state in the chain (i.e., the state that will play the role of q' in the computation of $a(q'|q)$). We assume a $\mathbb{Q}(q'|q)$ that restricts reachable states q' to those that differ from the current state q by the value of one binary variable. The mutation sampler derives its name from the fact that potential proposals are those formed by "mutating" the value of a single variable. Each mutated (reachable) state has an equal probability of being selected as a proposal. Edges in Figure 1B denote reachable states from some starting state. 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). In addition, later we present experimental results that provide direct empirical support for this proposal distribution.

Note that this proposal distribution confers additional efficiency benefits. Because only one variable V_i is changed,

the ratio $\frac{\pi(q')}{\pi(q)}$ simplifies to

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

where v_i and v_i' are the values of variable V_i in q and q', respectively, and u_i denotes the state of the variables in V_i's *Markov blanket*. By definition, variables outside V_i's Markov blanket are independent of V_i given u_i; thus, $\pi(v_i|v_{-i}) = \pi(v_i'|u_i)$ is entailed. In a causal graphical model, a variable's Markov blanket includes its direct parents, its direct children, and the other direct parents of its direct children (Koller & Friedman, 2009). It is often convenient to again express $\frac{\pi(q')}{\pi(q)}$ as the relative likelihood of two (now partial) network states by noting that $\frac{\pi(v_i'|u_i)}{\pi(v_i|u_i)} = \frac{\pi(v_i', u_i)\pi(u_i)}{\pi(v_i, u_i)\pi(u_i)} = \frac{\pi(v_i', u_i)}{\pi(v_i, u_i)}$. Examples of variables' Markov blanket and the resulting simplification of $\frac{\pi(q')}{\pi(q)}$ are presented in Figure 2; quantitative examples of the calculation of MH transition probabilities are presented in Appendix A. In particular, Appendix A demonstrates how those transition probabilities reduce to simple expressions involving only the local probabilities that define how variables are generated from their parents (known as a graph's *conditional probability distributions*, or CPDs). The fact that CPDs are assumed to be explicitly represented as part of a causal graphical model means that a chain of samples can be generated with exceptional efficiency.[1] The General Discussion will consider additional efficiencies that can be realized depending on the particular causal judgment a reasoner is faced with.



Mutated variable	Markov blanket	Metropolis-Hastings ratio
E	C	$\pi(ce')/\pi(ce)$
A	B, C	$\pi(a'bc)/\pi(abc)$
C	A, B, E	$\pi(abc'e)/\pi(abce)$

FIGURE 2 Examples of the calculation of Metropolis-Hastings ratio. When the mutated variable is E, $\pi(q')/\pi(q)$ reduces to $\pi(ce')/\pi(ce)$ because E's Markov blanket consists of only C. That is, for purposes of this calculation the states of A, B, and D in q and q' can be ignored. When the mutated variable is A, the calculation of $\pi(q')/\pi(q)$ can ignore the state of D and E. When the mutated variable is C, it can ignore the state of D.

[1] Although the mutation sampler appears superficially similar to a Gibbs sampler, it differs because a single Gibbs update, formed by sequentially applying elementary Gibbs updates for each variable in the network, can produce a change to more than one variable. Although a variant of a Gibbs sampler can be defined in which each update is itself elementary (so that only one variable changes), the transition probabilities this sampler defines, $a(q'|q) = \pi(v_i'|u_i)$, are distinct from those of the mutation sampler, $a(q'|q) = min(1, \pi(v_i'|u_i)/\pi(v_i|u_i))$.

2.2 | Biased Starting Points

The model thus far is simply an efficient MH sampler for estimating a causal graph's joint distribution. Importantly, however, we introduce a bias in the starting point for sampling: Sampling always starts from one of the 'prototype' states, those in which nodes are either all 0 or all 1. For example, for the network in Figure 1A, the prototypes are the bottom left and top right corners of Figure 1B, states referred to as $y_A^0 y_B^0 x^0$ and $y_A^1 y_B^1 x^1$, respectively.[2] We suggest that prototypes readily come to mind as plausible states at which to start sampling because, if one ignores the details of the causal graph such as the strength, direction, or functional form of the causal relations, they are likely to be viewed as having relatively high joint probability. This is so because it is easy to ascertain that they are qualitatively consistent with the generative causal relations: there are no instances of a cause being present and an effect absent or vice versa. In fact, Appendix B demonstrates that this assumption is not only psychologically plausible, it often leads to more accurate causal inferences for a given sample size. This finding reflects the fact that the prototypes often *are* the high probability states of a causal system with generative causal links and so the chain of samples will converge to the true joint distribution more quickly if sampling starts with them as compared to the other network states. Note that the General Discussion will consider generalizations of the mutation sampler to causal graphs with inhibitory causal links. There we will consider the appropriate points to start sampling from such graphs.

We also propose that, depending on the domain being reasoned about, one of the two prototype states may come to mind more readily than the other. For example, some of the empirical studies we analyze later taught participants novel categories with inter-feature causal relations and explicitly informed them that some category features are more likely than others. In such cases, we introduce a *bias* parameter that adjusts the probability of initializing the chain at each prototype. Unless otherwise mentioned, *bias* is set to .50, meaning that the chain is equally likely to be initialized at either prototype state.

2.3 | Limited Capacity

Regardless of our proposal distribution and biased initialization, the chain of samples generated by the mutation sampler is guaranteed to converge to the normative joint distribution as defined by the causal graphical model. However, convergence is likely only when the number of samples is large. In contrast, we assume that people are resource-constrained and thus can only take a few samples (on the order of a dozen rather than thousands or millions). Following Bramley et al. (2017), we assume that people have a fixed *capacity* for sampling but vary in the number of samples taken for any particular judgment, albeit with the constraint that at least two samples are drawn (the initial prototype and one more). To instantiate these constraints, for each judgment we draw from a Poisson distribution with mean λ' a quantity k'; the number of samples taken is $k = k' + 2$ (and so the mean number of samples is $\lambda = \lambda' + 2$). Larger λ values signify that a participant has a capacity to take many samples and so would behave more in line with a normative causal graphical model. Smaller λ values signify a limited capacity to take samples and thus a stronger divergence from the normative model. In particular, when λ is small the mutation sampler will not have time to fully explore the state space and so will overestimate the probability of states near the starting point and underestimate the remaining states.

These effects are illustrated in Figure 3, which presents the joint distributions derived by the mutation sampler for two types of graphs: a common effect graph (panel A) and a common cause graph in which X is a cause of Y_A and Y_B (panel B). The blue line represents the normative joint distribution—that is, for each of the eight possible graph states, the (joint) probability of that state—generated by the causal graphical model under a particular parameterization (computed

[2] Throughout this article we use lowercase and superscripts to denote the state of a variable, that is, $v_i^j \equiv V_i = j$. Thus, x^0 denotes that X is absent, y_A^1 that Y_A is present, and so forth.

via the standard methods defined in Appendix A).[3] The red lines represent approximations of that distribution derived from the mutation sampler for three different values of λ: 4, 8, and 32. To make the predictions of the mutation sampler comparable to the normative model (and each other), the samples it generates have been normalized by dividing the number of visits to each state by the total number of samples.

A comparison of the joint distributions in Figure 3 sheds insight into how the mutation sampler works. For both graphs, the joint probabilities for the prototype states estimated by the mutation sampler are greater than those derived from the normative model, a consequence of sampling beginning at those states. Conversely, the mutation sampler's joint probabilities for the remaining network states are less than those of the normative model. We argue that the fact that mutation sampling reproduces the general shape of the true joint distribution explains why people draw approximately veridical causal inferences whereas the systematic deviations from that joint explains the common errors they make. Note that the magnitude of those deviations vary with the average sample size λ. When λ equals 32 the mutation sampler's joint distribution begins to approach that of the normative model, confirming that the mutation sampler converges to the normative model when the resources required for extensive sampling are available.

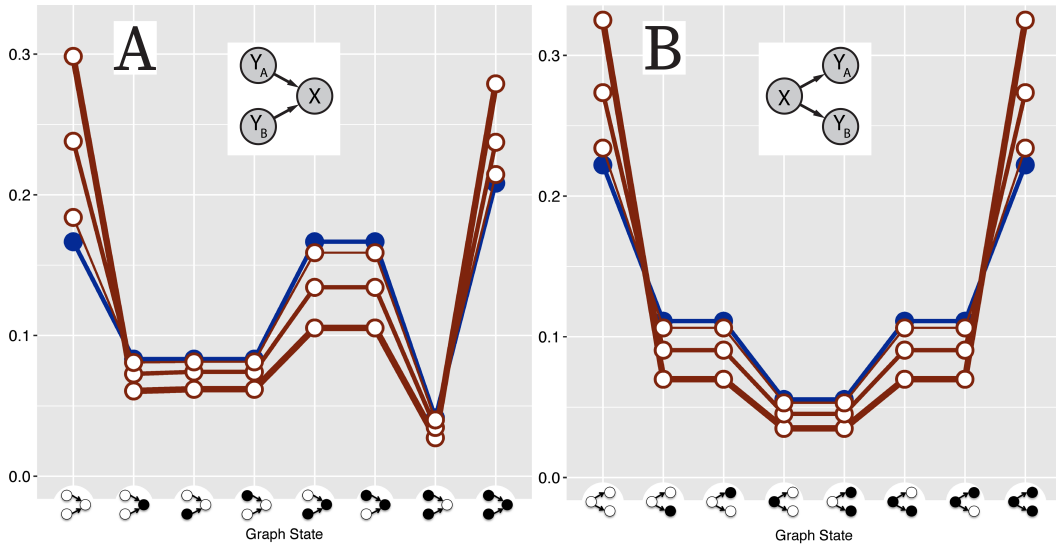


FIGURE 3 Joint probability distributions for (A) a common effect graph and (B) a common cause graph. The horizontal axis presents graph states in the same format as Figure 1 (i.e., filled circles indicate a 1, open circles a 0). Causal relations in both networks are assumed to be generative, independent, and combine according to a noisy-or integration function (see Appendix A). Both networks are parameterized such that the marginal probability of the causes = .50, the causal strength = .50, and the strengths of background causes = .33. The blue line (solid plot points) represents the joint distribution entailed by the normative model. Red lines (open plot points) represent the joint distributions implied by the mutation sampler, with thicker lines meaning fewer samples (thick: $\lambda = 4$; medium: $\lambda = 8$; thin: $\lambda = 32$).

[3] The probabilities in Figure 3 match what states intuition indicates should be more or less likely in light of the causal relations. For both causal graphs the prototype states in which variables are either all absent ($y_A^0 y_B^0 x^0$) or all present ($y_A^1 y_B^1 x^1$), shown on the far left and right, respectively, of each panel, are highly probable states. For the common effect graph, states in which the effect X is accompanied by one cause ($y_A^1 y_B^0 x^1$ and $y_A^0 y_B^1 x^1$) are moderately probable whereas the state where both causes are present but the effect absent ($y_A^1 y_B^1 x^0$) is quite improbable. For the common cause graph, states in which the cause X is accompanied by one effect ($x^1 y_A^0 y_B^1$) and ($x^1 y_A^1 y_B^0$) are moderately probable whereas one where both effects are absent ($x^1 y_A^0 y_B^0$) is not.

Also note that the mutation sampler's predictions in Figure 3 represent the joint distributions' *expected* values rather than a single run of the sampler. These expected values can be computed analytically. First, the distribution over the graph states representing which are likely to be the current state at sample n can be computed by multiplying the distribution at sample $n-1$ by the matrix of transition probabilities between graph states defined by the Metropolis-Hastings rule (the initial distribution before sampling commences is .50 at the two prototypes and 0 otherwise). Summing these distributions yields the expected number of visits to each graph state after n samples. Then, the expected joint associated with a given λ can be computed by taking a weighted average of the distributions computed at every chain length.[4] Nevertheless, remember that our psychological claim is that reasoners run a single chain of samples for each causal judgment. We will consider some consequences of that claim (e.g., the inherent variability of causal judgments) in the General Discussion.

2.4 | Example Run

The previous section presented the full formalization of how the mutation sampler generates a sequence of samples. However, it may be helpful to see the process at work. Table 1 shows a single run of the mutation sampler for the common effect graph in Figure 1A. The sample run shows every step of the process. The chain starts at one of the prototype states, in this case $y_A^1 y_B^1 x^1$ (referred to as 111 in the table for clarity). Then, a proposal that differs by only one variable ($y_A^1 y_B^0 x^1$, or 101) is generated. The MH ratio (the ratio of the proposal state's probability to the current state's probability) is calculated and compared to a random number generated from $Unif(0,1)$ (refer again to Appendix A for the calculation of MH ratios). If the ratio is greater than the random number, the current state is updated to the proposal state. If it is smaller, the proposal is rejected and the current state is unchanged.[5] As mentioned, if the size of the sample is very large it can be used to generate an estimate of the joint distribution that is indistinguishable from that specified by the normative model. If the size of the sample is limited, it will exhibit the sorts of deviations shown in Figure 3.

3 | EMPIRICAL TESTS

We now compare the mutation sampler to human judgments. Recall that our claim is that, when forming those judgments, people first generate a sample of concrete network states via a process of mutation sampling. This sample is then used as the basis for responses to a variety of causal judgment types. A key test of our model, then, is whether the biased samples generated by the mutation sampler underlie many different types of causal judgments. In particular, we compare fits of the normative model and mutation sampler on existing empirical datasets in reasoning, categorization, and intervention. To further test the task-generality of our model, we compare it to the normative model on a new task that directly asks for judgments of joint probability.

The general structure of this section is to push most of the specifics about fitting procedures to appendices, providing only high-level summaries of the relative performance of the mutation sampler and normative models. This allows us to put a spotlight on illustrative examples showing *why* the mutation sampler achieves better fits than the normative model, while still allowing motivated readers to delve into the particulars if they choose.

[4] Because in a Poisson distribution the probability of any value of k' is non-zero, we clipped that distribution so as to consider the values of k' that accounted for .999 of the distribution. For example, the values of k' from 0 to 15 account for .999 of a Poisson distribution with a mean of 8. We thus computed the 16 joint distributions associated with the values of k' of 0, 1, ..., 15 and then averaged those joints, each weighted by their probability (as specified by the Poisson distribution).

[5] To avoid division by zero in cases where a state was never visited, we initialize the number of visits to each state with a very small value (1e-10).

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 (marginal probability of causes = .50, causal strength = .50, and strength of background causes = .33).

While the mutation sampler was designed to model both the successes and failures of people's causal reasoning judgments, the particular cases discussed in this section mostly involve deviations from the normative model. The reason for this is simple: to understand how the mutation sampler achieves better fits than the normative model, we focus on aspects of behavior that the normative model cannot account for.

3.1 | 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 D presents fits of the normative model and the mutation sampler to 18 experimental conditions from four past articles, involving a total of 672 participants and 8 distinct causal network topologies. In every condition participants were first instructed on causal knowledge and then presented with a series of conditional probability queries, judgments in which participants estimate the probability that one variable is present given the state of one or more of the other variables.

For each participant, we fit versions of the normative model and the mutation sampler suitable for the causal network topology taught to that participant. For both models, those parameters included one representing the marginal probability of the variables that are root causes in a graph (e.g. Y_A and Y_B in Figure 1), one representing the strength of 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 scaled that participants us. The mutation sampler had an additional λ parameter representing mean chain length, that is, the mean number of samples taken. Details of the fitting procedure and the best fitting average parameter values are presented in Appendix D for each condition along with a number of measures of quality of fit.

Appendix D 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 participants 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 consider the three-variable common cause condition ($Y_A \leftarrow X \rightarrow Y_B$) in Figure 4A. The Markov condition associated with causal graphical models and alluded to earlier stipulates that a variable is statistically independent of its non-descendents conditioned on the state of it immediate parents. Applied to the common cause graph, the Markov condition states that the two Ys should be independent conditioned on X such that the probability of one Y (call it Y_i) given the state of X should be unaffected by whether the state of the other Y (Y_j) is present, absent, or unknown. In other words, it should hold that $p(y_i^1 | x^0 y_j^0) = p(y_i^1 | x^0) = p(y_i^1 | x^0 y_j^1)$ and $p(y_i^1 | x^1 y_j^0) = p(y_i^1 | x^1) = p(y_i^1 | x^1 y_j^1)$.[6] These predictions are represented by the two horizontal blue lines in Figure 4A, which depicts the normative model's best fit to these data. The figure reveals that participants instead violated the conditional independence required by the Markov condition, judging that, for example, $p(y_i^1 | x^1 y_j^0) < p(y_i^1 | x^1) < p(y_i^1 | x^1 y_j^1)$. This finding has been replicated 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; 2018; Rottman & Hastie, 2016; Walsh & Sloman, 2004; see Hagmayer, 2016, and Rottman & Hastie, 2014, for reviews). Figure 4A also reveals that those independence violations are reproduced by the mutation sampler. The independence violations predicted by the mutation sampler are a direct consequence of its biased starting points combined with a relatively small number of samples.

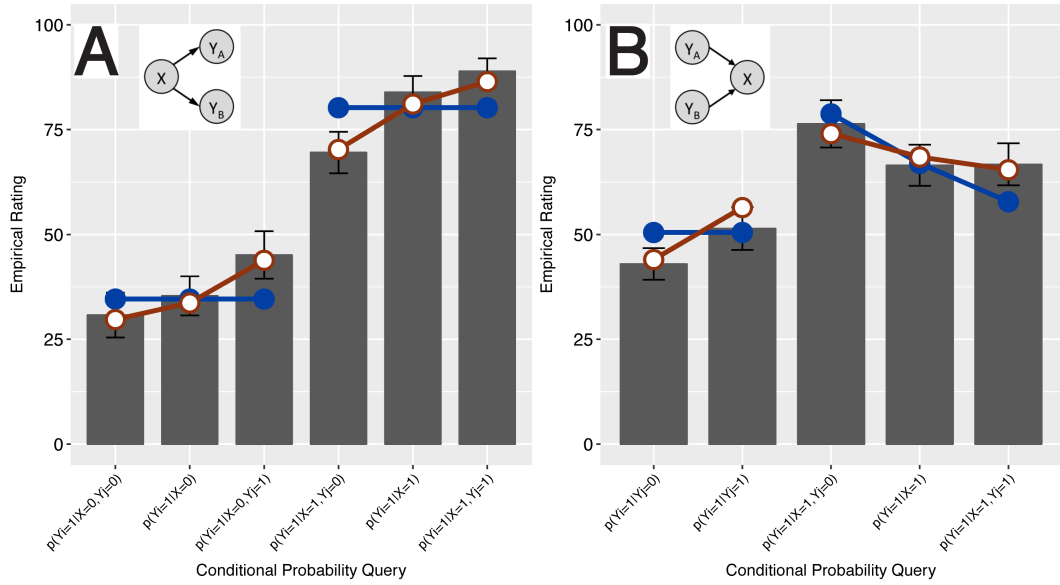


FIGURE 4 Data from Rehder & Waldmann (2016), Experiment 1. Fits of the mutation sampler (red lines, open plot points) and the normative model (blue lines, solid plot points) are superimposed on the empirical conditional probability judgments (gray bars). For example, in these studies participants judged the presence of the to-be-inferred variable on a 0-100 scale. Error bars denote 95% confidence intervals.

Second, for the three-variable common effect condition ($Y_A \rightarrow X \leftarrow Y_B$; Figure 4B), the normative model stipulates

[6] Because of the symmetry of the two causal networks in Figure 4 and the study's extensive counterbalancing of materials, we generally collapse over inferences involving variables that play interchangeable roles, such as Y_A and Y_B in Figure 4. For example, the rating for the conditional probability judgment $p(y_i^1 | x^1)$ shown in the figure is the average of $p(y_A^1 | x^1)$ and $p(y_B^1 | x^1)$, $p(y_i^1 | x^1 y_j^0)$ is the average of $p(y_A^1 | x^1 y_B^0)$ and $p(y_B^1 | x^1 y_A^0)$, and so forth.

that the two Ys should be unconditionally independent. That is, it should hold that $p(y_i^1 \mid y_j^0) = p(y_i^1 \mid y_j^1)$. This prediction is represented by the horizontal blue line in Figure 4B, the normative model's best fit to these data. Yet participants judged that $p(y_i^1 \mid y_j^0) < p(y_i^1 \mid y_j^1)$ 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; 2018; Rottman & Hastie, 2016; Trueblood et al., 2017; cf. Von Sydow et al., 2010). Figure 4B also reveals that the mutation sampler correctly accounts for this violation of independence.

The mutation sampler also accounts for another reasoning error that participants commit with the common effect graph in Figure 4B. Explaining away is a signature property of common effect graphs with independent generative causes. If X is observed to occur then the probability that Y_i is present of course increases. But if it is then further observed that the other cause Y_j is present then the probability that Y_i is present should decrease back towards its baseline. Conversely, if Y_j is observed to be absent then the probability of Y_i should increase. That is, it should hold that $p(y_i^1 \mid x^1 y_j^0) > p(y_i^1 \mid x^1) > p(y_i^1 \mid x^1 y_j^1)$. In fact however, research finds that participants often explain away too little or not at all (Morris & Larrick, 1995; Rehder, 2014a; 2018; see Rottman & Hastie, 2014, for a review). The right three bars in Figure 4B illustrate the three conditional probability judgments relevant to explaining away. The fits of the normative model to these data points reveal that explaining away with Rehder and Waldmann's participants 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).

Finally, one of the conditional reasoning studies presented in Appendix D allows an assessment of the mutation sampler's proposal distribution in which a proposal state is allowed to differ from the current state by one variable. By testing causal networks with five variables — an extended common cause network ($Z_A \leftarrow Y_A \leftarrow X \rightarrow Y_B \rightarrow Z_B$) and an extended common effect network ($Z_A \rightarrow Y_A \rightarrow X \leftarrow Y_B \leftarrow Z_B$) — Rehder (2018) found that reasoners violated independence on a number of different types of conditional probability queries. Appendix C demonstrates that whereas the mutation sampler is able to account for each of those types, an alternative sampler in which proposal states may differ from the current state by more than one variable is not. The conditional probability judgments on which the alternative sampler fails are those in which the antecedent involves a network state with mixed 1s and 0s. Independence violations on such problems obtain only when the bias introduced by starting sampling at the prototype states is allowed to percolate through the network towards such mixed states gradually. Quantitative examples of this process are presented in Appendix C.

3.2 | Causal Categorization

We also assess how the mutation sampler accounts for categorization judgments drawn on the basis of causal knowledge. Appendix E presents fits of the normative model and the mutation sampler to 25 experimental conditions from 7 published articles, involving a total of 1044 participants and 9 distinct causal network topologies. In every condition participants were instructed on novel categories whose features were causally related. For example, some participants were informed of a type of star named Myastars with binary dimensions such as type of helium (ionized or not), density (high or normal), number of planets (large or normal), and so forth. The generative causal relations were described as one feature causing another (e.g., ionized helium causes a large number of planets). After learning their assigned category participants were presented with test items with features and asked to rate on a 0-100 scale the likelihood that each was a member of the category.

We fit versions of the normative model and the mutation sampler suitable for the causal network topology in each experimental condition. Because the materials were categories, their description included information specifying that one feature on each binary dimension was more prevalent than the other. For example, in one study, participants were

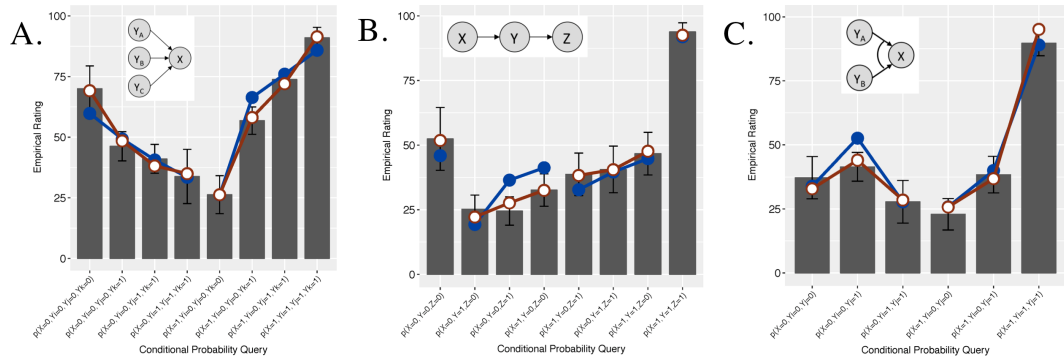


FIGURE 5 Categorization data from (A) Rehder (2003a), common effect condition; (B) Rehder and Kim (2010), Experiment 1, weak condition; and (C) Rehder (2014b), Experiment 1, conjunctive condition. Fits of the mutation sampler (red lines, open plot points) and the normative model (blue lines, solid plot points) are presented superimposed on the empirical data (gray bars). Error bars denote 95% confidence intervals.

told that "Most Myastars have high density whereas some have normal density." Because of these uneven base rates, we granted the mutation parameter the additional *bias* parameter described above that controls the probability that sampling process starts at the prototype state with all 1s versus the one with all 0s.

Details of the studies and fitting procedure are presented in Appendix E, along with the best fitting parameters and measures of fit. It reveals that the mutation sampler yielded a better fit (correcting for the number of parameters) as compared to the normative model in 21 of the 25 experimental conditions. Figure 5 presents categorization queries from three of those conditions, which again superimposes the fits of the normative model and mutation sampler on the empirical ratings. Note that the values of the fitted bias parameter was greater than 0.5 in the large majority of studies, suggesting that it indeed reflected the fact that one feature on each dimension was considered characteristic of the category and the other was uncharacteristic.

Figure 5A presents the categorization ratings associated with a common effect network with three cause features. The eight distinct types of categorization test items presented in the figure are organized into two groups of four. The group on the left includes the test items in which the common effect X is absent. As the number of Ys that are present increases from zero to three, categorization ratings decrease, because more Ys means more violations of the causal relations. The group on the right includes the test items in which X is present. For these items ratings increase with more Ys because more causal relations are confirmed (and, generally, because more characteristic features makes for a better category member). Although the normative model (blue lines) accounts for this qualitative pattern, it underestimates the joint probability of the two prototypes ($y_A^0 y_B^0 y_C^0 x^0$ and $y_A^1 y_B^1 y_C^1 x^1$; see outer bars in the figure) and slightly overestimates the states in between. In contrast, the mutation sampler (red lines) yields a better account of these data, including the two prototypes.

Figure 5B presents results from a three feature causal chain. From left to right, the eight test items are organized into (a) the $x^0 y^0 z^0$ prototype, (b) the three items in which one feature is present, (c) the three in which two features are present, and (d) the $x^1 y^1 z^1$ prototype. Within both the one-feature and two-feature items, the items on the left ($x^0 y^1 z^0$ and $x^1 y^0 z^1$) violate two causal relations whereas the other members of each group violate only one. Although participants' ratings reflect this difference in the number of violated causal relations, the panel makes clear that they do not do so to the degree predicted by the normative model. In contrast, the mutation sampler reproduces these ratings

because it reduces differences in the joint probability of non-prototype items.

Figure 5C presents results from a common effect network with two features that are a conjunctive cause of a third. The six types of test items are organized into two groups. X is absent in the group on the left and present in the one on the right. When X is absent, the normative model predicts that the joint probabilities will increase with the introduction of the first Y cause feature but then sharply decrease with the introduction of the second, because when both Ys are present the absence of X represents a violation of the (conjunctive) causal relation. Participants ratings reflected this pattern, but less sharply than the normative model. The mutation sampler, in contrast, reproduced participants' ratings.

3.3 | Causal Interventions

The third type of task we account for with the mutation sampler is causal interventions. Whereas in a typical causal inference a reasoner observes the states of one or more variables and then predicts the state of another variable, in an intervention the reasoner is asked to imagine that an agent external to the causal system has acted on the system so as to set one of the variables to a particular state. Causal graphical models stipulate that valid inferences depend on whether variables are observed or intervened upon.

For example, consider the two causal models tested by Waldmann and Hagmayer (2005, Experiment 1) shown above Figure 6A. Participants were told they were being instructed on a sleeping sickness in which a mosquito bite causes the production of a substance named pixin. In one condition, participants were taught the causal relations between *pixin (P)* and *xanthan (X)*, *sonin (S)*, *gastran (G)*, and *histamine (H)* depicted by Model A in Figure 6A. In another they were taught Model B, which is identical to Model A except that the direction of the causal relationship between X and G is reversed. All participants then observed training data consisting of 20 patients and their values on each of the five variables. This data reflected deterministic causal relations (a cause was always accompanied by its effect) and a base rate for P of 0.5. Participants then predicted the state of S given the state of H in a particular patient. Whereas in the observation condition participants merely observed H to be high (or low), in the intervention condition they were told that a doctor had inoculated the patient with a substance that raises (or lowers) the level of H (such interventions are depicted by a double-lined arrow in Figure 6A and are denoted $do(h^1)$). The import of the state of H being due to an intervention is that an inference from H to P is no longer licensed: according to the logic of graph surgery (Pearl, 1988), an intervention on H can be modeled by the removal of the $P \rightarrow H$ causal relation (depicted by a dashed line in Figure 6A). Thus, the difference between the observation and intervention condition is that whereas H provides information about S in either condition in Model B (via the path $H - G - X - S$), it does so only in the observation condition in Model A (the $H - P - X - S$ path is blocked when H is intervened upon).

The empirical results from Waldmann and Hagmayer's (2005) Experiment 1 are shown in Figure 6A. The four inferences with Model A (H = high vs. low crossed with H observed or intervened upon) are shown on the left and those with Model B are shown on the right. For Waldmann and Hagmayer's purpose, the important finding was that participants judged H to be highly diagnostic of S in Model B regardless of whether H was observed or intervened upon, whereas for model A H was regarded as diagnostic of S when H was observed, but not when it was intervened upon. This result highlighted the fact that reasoners are in fact quite sensitive to the different inferences that are licensed when a variable is intervened upon rather than observed. For our purpose, the important finding is that whereas the difference between $p(s^1 \mid do(h^1))$ and $p(s^1 \mid do(h^0))$ in Model A was indeed modest, it was clearly greater than zero. This is the case despite the fact that H and S are conditionally independent in Model A after the surgery that removes $P \rightarrow H$. That is, from the perspective of the normative causal graphical model framework, the results in Figure 6A represent another form of independence violation.

To illustrate the performance of the mutation sampler in this context, we fit it (red line) and the normative model

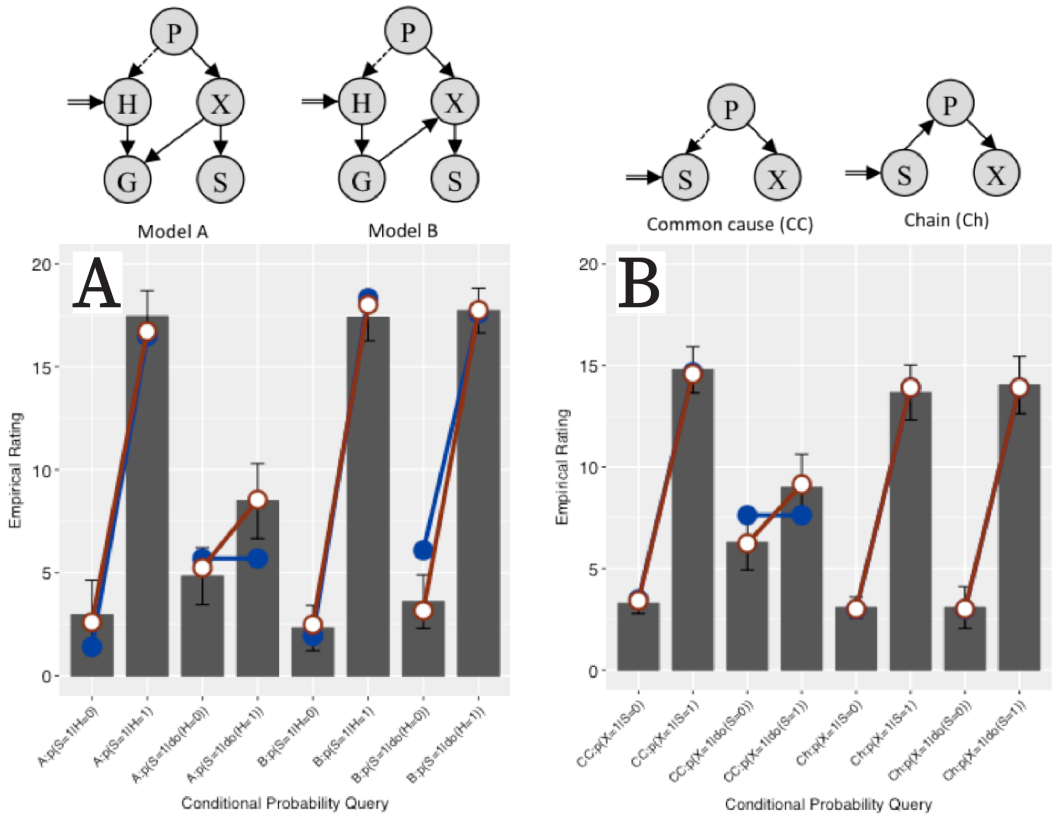


FIGURE 6 Intervention data from Waldmann and Hagmayer (2005), Experiments 1 (panel A) and 2 (panel B). Double-lined arrows indicate interventions.

(blue line) to the aggregated ratings in each experimental condition. Figure 6A shows that both models account for most data points. But whereas the normative model is constrained to predict $p(s^1|do(h^1)) = p(s^1|do(h^0))$, the mutation sampler correctly predicts that $p(s^1|do(h^1)) > p(s^1|do(h^0))$.[7]

Figure 6B presents two other experiments reported by Waldmann and Hagmayer (2005). As was the case for Models A and B in Figure 6A, the common cause and chain models license the same qualitative inferences under observation but not under intervention (in this experiment, variable S was observed or intervened upon and X was the to-be-predicted variable). In particular, whereas the normative common cause model predicts $p(x^1|s^1) > p(x^1|s^0)$, it also predicts $p(x^1|do(s^1)) = p(x^1|do(s^0))$ because of the surgery that removes $P \rightarrow S$ causal relation. But although participants' inferences again showed that they were quite sensitive to observations versus interventions, they incorrectly judged that $p(x^1|do(s^1)) > p(x^1|do(s^0))$. The fits in Figure 6B superimposed on the empirical ratings show that this independence violation can be accounted for by the mutation sampler but not the normative model. Note

[7] Note that because participants were only asked a relatively small number of conditional probability judgments in this study, we do not present quantitative measures of the fit of the two models as we did in the previous sections. Indeed the excellent quantitative fit of the mutation sampler in Figure 6A is likely due to the overfitting. Rather, our purpose is to simply demonstrate that mutation sampler can account qualitatively for a violation of independence in the context of interventions that the normative model cannot.

that unlike the first experiment, the data accompanying the causal models implied causal links that were quite strong but not deterministic, demonstrating that independence violations in the context of interventions are not limited to deterministic relations.

4 | CAUSAL REPRESENTATIONS

One key aspect of the empirical results presented thus far is that the mutation sampler accounts for not only a large number of experimental conditions and participants but also a number of different types of causal judgments. This result lends support to our assumption that the samples generated by mutation sampling serve as the basis for many judgment types and that the small but systematic distortions relative to normative responses introduced by that sampling will therefore manifest themselves on multiple tasks. Our final empirical test of the mutation sampler involves a novel methodology that provides a relatively direct assessment of participants' beliefs about the relative likelihood of the states of a causal graph. We ask whether the distortions introduced by mutation sampling — distortions we have implicated as the source of the reasoning errors in the previous sections — can be observed directly in participants' own generated samples.

4.1 | Method

4.1.1 | Materials

Participants were presented with causal graphs 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"). For half of the participants the causal relations on which they were instructed formed common effect graph ($Y_A \rightarrow X \leftarrow Y_B$), whereas for the other half they formed a common cause graph ($Y_A \leftarrow X \rightarrow Y_B$). For example, in the domain of economics one of the causal relationships was "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." See Rehder (2014a) for additional examples of the causal relations.

4.1.2 | Procedure

Participants first studied screens of information that defined the variables, presented the mechanisms via which 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. Participants 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. Participants were asked to generate an analogous distribution in their assigned domain (economics, etc.). Specifically, they were given 50 U.S. 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).[8]

4.1.3 | 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-participants design. 120 New York University undergraduates received course credit for participation.[9]

4.2 | Results

Initial analyses revealed no effect for domain or variable senses, so results were collapsed over these factors. As expected, the allocation of coins differed depending on whether participants were instructed on a common cause or common effect graph and thus the results from these two conditions are presented separately.

4.2.1 | Common effect

Figure 7A presents how participants allocated the 50 coins to the eight states of a common effect graph (gray bars), states depicted in the original Figure 1A. First note that these allocations indicate that participants judged that the two prototype states ($y_A^0 y_A^0 x^0$ and $y_A^1 y_A^1 x^1$), shown on the far left and far right of Figure 7A, respectively, were the most probable whereas the rest of the states were less probable. This result was expected and indicates that participants attended to the causal relations they learned. However, the important theoretical question is whether their distributions of coins reflects the kinds of distortions relative to the normative model predicted by the mutation sampler. To answer this question, Figure 7A also presents the best fits of both the normative model (blue line) and the mutation sampler (red line) superimposed on the empirical data.[10] Note that, relative to the normative model, the mutation sampler overpredicts the number of coins for the two prototype states and underpredicts the remaining states, a pattern that of course reflects the theoretical predictions presented earlier in Figure 3A. More importantly, the figure makes clear that participants' distributions of coins were better accounted for by the mutation sampler. Indeed, a measure of fit that corrects for the difference in the number of parameters in the two models (*AIC*) confirmed that the mutation sampler yielded a better fit as compared to the normative model (3.6 vs. 10.8).

We also wanted to relate the results in Figure 7A to the findings reported earlier regarding how people answer conditional probability queries. To do so we derived measures that reflect the statistical relationships among the three variables implied by a participant's distribution of coins. In particular, we first normalized that distribution and then computed the phi coefficient between a Y and an X, $\phi(Y_i, X)$ (the coins 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 participants are presented in Figure 7B. First note that the fact that $\phi(Y_i, X) >> 0$ indicates that participants distributed the coins in a manner that reflected positive correlations

[8] In particular, for each of the three domains there were four alternate value configurations (all "High"; all "Low"; two "High" and one "Low"; and one "High" and two "Low"). For each of these twelve counterbalanced conditions, there were two separate sheets with the concrete cases randomly assigned (resulting in a total of 24 sheets).

[9] This study was approved by the New York University Institutional Review Board under protocol number IRB-FY2017-68.

[10] The best fitting parameters were those that maximized the likelihood of the normalized distribution of coins. Averaged over participants, those parameters were c = .519, m = 0.440, and b = .243 for the normative model and c = .534, m = 0.410, b = .328, and λ = 10.1 for the mutation sampler. To make them comparable to participants' distribution of coins, in Figure 7A the models' fits have been multiplied by 50.

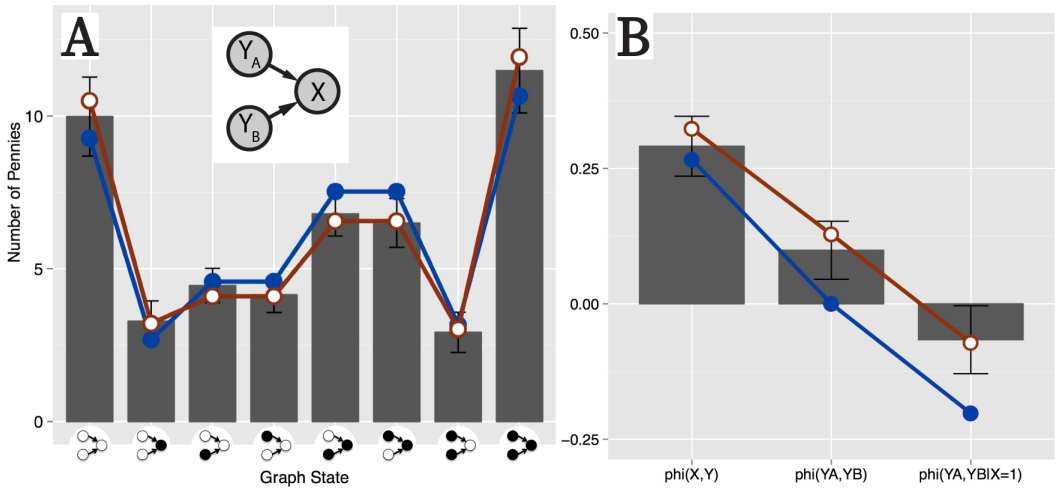


FIGURE 7 Results from the common effect condition. (A) Fits of the mutation sampler (red line, open plot points) and normative (blue line, solid plot points) are presented superimposed on participant's distribution of coins (gray bars). (B) Phi coefficients derived from participant's judgments in the common effect condition along with those derived from corresponding mutation sampler and normative model fits. Error bars denote 95% confidence intervals.

between the Ys and X and thus their understanding 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, $t(59) = 3.62$, $p < .001$. That is, the positive correlation between the causes of a common effect graph, observed earlier in people's causal inferences (see Figure 4B) and representing an independence violation, also manifested itself in their distribution of the coins. Finally, note that $\phi(Y_A, Y_B|x^1)$ is significantly less than 0, $t(59) = -2.07$, $p < .05$. That is, the negative correlation between the causes conditioned on the presence of the common effect observed earlier in people's explaining away inferences (Figure 4B) also manifested itself in their distribution of coins.

To directly assess the models' ability to account for these effects, we also derived the phi coefficients implied by their fits to the distributions of coins. These coefficients are superimposed on the empirical phi coefficients in Figure 7B. As expected, the normative model, which stipulates that the Ys are independent (i.e., that $\phi(Y_A, Y_B) = 0$), is unable to account for the fact $\phi(Y_A, Y_B) > 0$. And although the normative model correctly predicts that $\phi(Y_A, Y_B|x^1) < 0$, it sharply overestimates the magnitude of that effect. Once again, we see that the mutation sampler but not the normative model accounts for the independence violations and the too-weak explaining away exhibited by human reasoners.

4.2.2 | Common cause

Figure 8A presents how participants allocated the 50 coins to the eight states of a common cause graph. To again evaluate how the normative and mutation sampler models account for these results, Figure 8A shows their fits to these data.[11] Although the fits of the 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

[11] The best fitting parameters for the normative model were $c = .510$, $m = .556$, and $b = .285$. Those for the mutation sampler were $c = .472$, $m = .466$, $b = .323$, and $\lambda = 8.3$.

(*AIC*s of -23.0 vs. -14.1).

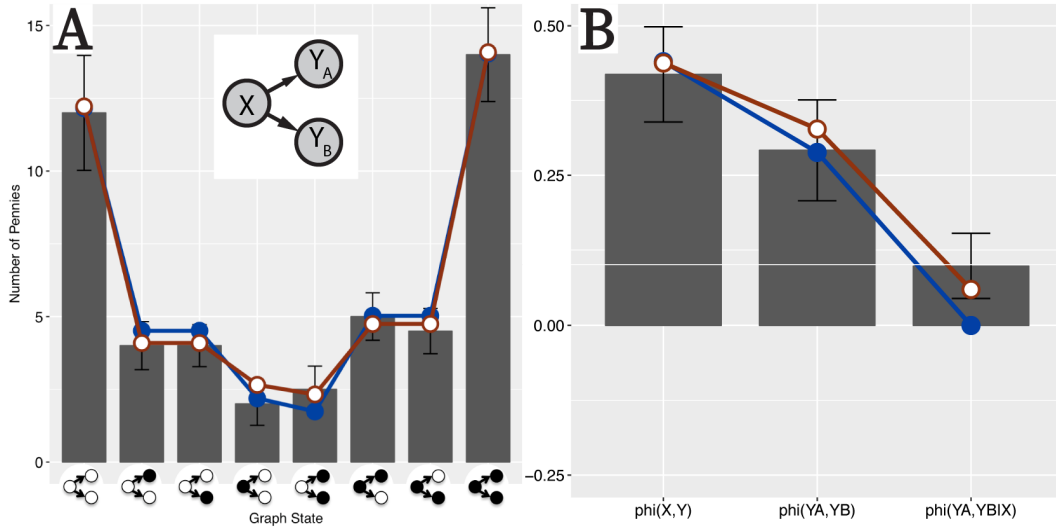


FIGURE 8 Results from the common cause condition. (A) Fits of the mutation sampler (red line, open plot points) and normative (blue line, solid plot points) are presented superimposed on participant's distribution of coins (gray bars). (B) Phi coefficients derived from participant's judgments in the common effect condition along with those derived from corresponding mutation sampler and normative model fits. Error bars denote 95% confidence intervals.

To relate these results to the earlier conditional reasoning studies with common cause networks, just as in the common effect condition we derived statistical measures characterizing the joint, which are presented averaged over participants in Figure 8B. That $\phi(X, Y_i) >> 0$ (i.e., that X was viewed as positively correlated with the Ys) indicates that participants understood that X was a generative cause of the Ys. That $\phi(Y_A, Y_B) >> 0$ (i.e., that the Ys were viewed as positively correlated with each other) indicates that participants correctly understood that, in a common cause graph, one Y predicts another. The final coefficient in Figure 8B, $\phi(Y_A, Y_B|X)$, represents the degree to which the Ys were viewed as positively correlated conditioned on X.[12] The Markov condition associated with causal graphical models of course stipulates that the Ys are independent conditioned on X, that is, that $\phi(Y_A, Y_B|X) = 0$. Participants in this condition judged instead that $\phi(Y_A, Y_B|X) > 0$, $t(59) = 3.60$, $p < .001$. That is, the violation of independence between two effects supposedly screened off by their common cause, observed earlier in people's causal inferences (Figure 4A), is also observed here in their distributions of the coins. The phi coefficients derived from the model fits in Figure 8B shows that the mutation sampler but not the normative model can account for the fact that $\phi(Y_A, Y_B|X) > 0$.

4.2.3 | Discussion

The results of this experiment confirm that the distortions introduced by mutation sampling—distortions we have implicated as the source of the errors that obtain during causal reasoning, categorization, and interventions—can be observed directly in participants' own generated samples. We do not wish to claim that the mental processes invoked

[12] $\phi(Y_A, Y_B|X)$ was computed as the average of the cases where X was present, $\phi(Y_A, Y_B|x^1)$, and absent, $\phi(Y_A, Y_B|x^0)$.

by a task that requires explicit construction of a causal sample corresponds exactly to those involved in these earlier judgments. Nonetheless, it is striking that the distributions that participants judged to be representative of causal graphs exhibited the same statistical properties—independence violations and weak explaining way—exhibited earlier. That the phenomena predicted by the mutation sampler manifest themselves on such a variety of judgments suggests that it has implications not just for specific tasks but for causal cognition more generally.

5 | GENERAL DISCUSSION

Although causal graphical models have enjoyed success in modeling causal cognition, less work has investigated the cognitive processes by which such sophisticated judgments are made. We have introduced a rational process model that generates samples from a causal system upon which many causal-based judged can be based. In fact, we found that the mutation sampler yielded a better fit across a diverse array of experimental conditions encompassing over 1,700 participants and confirmed a key prediction of the model on another 120 participants using a new methodology that assessed, in a relatively direct way, people's causal representations. In particular, we showed how the mutation sampler both reproduces humans' generally good causal reasoning performance while also accounting for the systematic errors they make, such as independence violations and weak explaining away. Nevertheless, as a rational process model, the mutation sampler embodies the view that humans *could* draw veridical causal inferences — if only they had the cognitive resources to do so. The fault thus lies in the fact that causal judgments must be computed in finite time and with limited resources. Errors in causal cognition are thus an unavoidable consequence of the tradeoff between accuracy, speed, and effort.

Below we first revisit the constraint of cognitive efficiency that any process model is obligated to satisfy. The mutation sampler is then compared to some alternative accounts of causal reasoning. We then consider some phenomena in causal cognition that are unaddressed by computational level models (such as causal graphical models) but potentially explainable by a process model such as the mutation sampler, including the variability in causal judgments, individual differences, and reaction times. We close with a discussion of generalizations to the mutation sampler to causal scenarios that have received relatively less empirical study.

5.1 | Sampling and Efficiency

As mentioned, we believe that people generally exhibit an impressive ability to reason causally, a fact that explains why the causal graphical model framework has enjoyed success in modeling causal cognition. Any process model purported to account for this phenomenon is therefore constrained to postulate cognitive mechanisms whose resource demands are within reach of most human reasoners for most causal based judgments. For this reason we believe it is important to emphasize again that the mutation sampler constructs samples in a manner that is computationally efficient and so psychologically plausible. The Metropolis-Hastings rule combined with the proposal distribution we advocate requires computing the relative likelihood of two graph states that differ by one variable, excluding variables not in the variable's Markov blanket. As demonstrated in Appendix A, this computation can be reduced to simple expressions involving the probabilities that define how variables are generated from their parents, probabilities that are assumed to already be explicitly represented as part of a causal graphical model.

Another aspect of efficiency of course concerns the size of the representations that are required. A well-known advantage of causal graphical models is that, by only encoding the local dependencies between variables and their parents, they avoid the exponential explosion in the space required if causal systems were represented as full joint

distributions. Of course, the purpose of the mutation sampler is to approximate that joint distribution, which raises the question of whether drawing inferences via sampling reintroduces the problem of unrealistic space requirements that causal graphical models were intended to solve in the first place. However, note that sampling needs to represent not all network states but rather only those that are visited (i.e., sampled). For example, if one runs the mutation sampler on the five variable graph of Figure 2 (which has $2^5 = 32$ distinct states) with chain lengths of 6, 12, 24, and 36 under the same parameterization specified in Appendix A, the average number of distinct network states that are actually sampled are 2.9, 4.9, 7.8, and 10.0, respectively. This exercise reveals that a psychologically plausible amount of sampling is accompanied by psychologically plausible space requirements. Later we will provide evidence that the small number of sampled network states results in variability that qualitatively matches that of participants.

Additional efficiencies are possible depending on the type of judgment a reasoner is faced with. Although the approach in this article has been to use the mutation sampler to estimate a joint distribution, which is then used to derive predictions for a specific task, the full joint distribution is often unnecessary. For example, when estimating a conditional probability, sampling the part of a causal network's state space in which the query's antecedent is false is a waste of cognitive resources. To investigate this opportunity for further optimization, we defined an alternative version of the mutation sampler. Whereas in the mutation sampler's proposal distribution each mutated state has an equal chance of being proposed, in the alternative sampler's proposal distribution those network states in which a conditional probability query's antecedent is true are ten times more likely to be proposed than those in which it is false. This modification means that sampling is strongly biased towards that part of the network's state space needed for the computation of that conditional probability. Appendix F confirms that, for the network in Figure 2, the rate at which conditional probability queries estimated via sampling converge to the true conditional probabilities is faster for this alternative sampler as compared to the standard mutation sampler. Indeed, for those particular queries the average accuracy achieved by the mutation sampler in 12 samples was matched by the alternative sampler after only 5.8 samples. Of course, computing conditional probabilities via sampling is also highly space efficient, because one need not remember visited network states at all. Rather, one only need keep a tally of the number of visited states that satisfy the antecedent and, of those, the number that satisfy the consequent (and then divide the latter by the former).

5.2 | Alternative Models

5.2.1 | Beta-Q

One model that shares some important similarities with the mutation sampler is Rehder's (2018) *beta-Q* model. Like the mutation sampler, beta-Q proposes that people draw inferences on the basis of non-normative joint distributions and, moreover, that it is the homogeneous "prototype" states that are overrepresented in those distributions. That the joints defined by the two models share these properties of course means that they make many of the same predictions. For example, both models not only predict the basic independence violations and weak explaining away results documented by numerous investigators (e.g., those shown earlier in Figure 4), but also some newer phenomena reported in Rehder (2018), such as the fact that independence violations arise even when causal inferences are screened off by three variables. In fact, the data sets from Rehder (2018) comprised six of the causal reasoning conditions successfully fit by the mutation sampler and one particular result from that study was cited as support for the mutation sampler's proposal distribution (see Appendix D).

However, an important difference between the models is that whereas beta-Q is a descriptive account, we advocate the mutation sampler as a process level account of the mental operations that underlie causal judgments. For example, the distortions to a normative joint distribution stipulated by beta-Q were defined in terms of *energy functions* that

were, in essence, added to a normative joint in order to achieve the needed distortion. But although beta-Q replicated participants' behavior, it provides no explanation of why joint distributions are distorted in that manner (or at all). In contrast, the mutation sampler provides an explanation for why those distributions might arise, namely, as a result of a mental sampling process that is limited and constrained to commence at certain easily imagined network states. For these reasons, we believe that the mutation sampler can be viewed as the process level implementation of the computations specified by beta-Q.[13] Of course, as a process model the mutation sampler has the potential to also account for some of the behaviors we discuss below (e.g., reaction times) that are outside the purview of descriptive models like beta-Q.

5.2.2 | Quantum Probability Models

Recently, Trueblood, Yearsley and Pothos (2017) applied *quantum probability theory* (QP) to some of the causal reasoning phenomena considered here. Whereas in classic probability theory probabilities are computed as subsets (of a joint probability distribution), in QP events are represented as subspaces and probabilities are computed by taking the inner product of vectors within those spaces. One key factor that determines QP probabilities is the *dimensionality* of the representational space deemed appropriate for a domain. When all N (binary) domain variables are *compatible*, then the dimensionality of the space is 2^N and QP can yield classic probabilities. But *incompatible* variable pairs result in a reduction in dimensionality (to 2 in the extreme case in which all pairs are deemed incompatible). This reduction addresses the exponential growth in space requirements alluded to above but at the cost of introducing so-called *quantum effects*, probabilities that do not necessarily honor the properties of classic probability theory. For example, QP interprets conjunctive probabilities in terms of a sequence of vector operations. Because $p(X, Y)$ and $p(Y, X)$ involve distinct sequences of such operations, in low dimensionality spaces commutativity in which $p(X, Y) = p(Y, X)$ need not hold.

One strength of QP is that it provides a natural account of order effects. For example, if $p(Z|X, Y)$ is interpreted as $p(Z|X then Y)$ (i.e., one first learns that X and then that Y), then QP potentially accounts for the finding that the order of X and Y matters (i.e., that $p(Z|X then Y) \neq p(Z|Y then X)$. In fact, Trueblood et al., (2017) found order effects in a causal reasoning task such that more recently presented information was weighed more heavily (also see Trueblood & Busemeyer, 2012). They also found that appropriately parameterized QP models could reproduce Markov violations and weak explaining away, phenomena we have taken here as evidence for the mutation sampler.

Although QP is a potentially important contribution to the understanding of human probabilistic reasoning, as presently formulated its application to causal reasoning is incomplete. Thus far, QP has only been applied to one causal network topology (a three-variable common effect network, Trueblood & Busemeyer, 2012; Trueblood et al., 2017). More generally, no principles have been specified that determine which variables in a causal network should be treated as compatible or incompatible (and thus the dimensionality of the space) as a function of their causal roles. As a result, QP makes no a priori predictions regarding how, for example, judgments should differ between a common effect and a common cause network. The asymmetries in human judgments elicited by these two network topologies have provided key evidence in favor of the causal graphical model framework. Similarly, within a QP space no principles guide the values of free *rotation parameters* (that determine the relations between the basis vectors that represent incompatible variables) as a function of, say, the strengths of the causal relations (and thus QP makes no a priori predictions that stronger causal relations should support stronger inferences). Finally, and most importantly for present purposes, QP is

[13] The beta-Q model includes a parameter q that determines the degree of distortion to the joint distribution, where larger distortions are implied by large values of q (and no distortion is implied when $q = 0$). Thus, q is inversely related to the chain length defined by the mutation sampler (where longer chain lengths imply more veridical joint distributions). In fact, we have established that chain lengths in the range [2, 7] yield causal inferences that are virtually identical to those generated by beta-Q with q in the range [1.33, 0.50]

fundamentally a descriptive model of the causal reasoning phenomena investigated here. Just like beta-Q, QP specifies computations that are intended to reproduce human causal inferences without specifying how those computations are carried out. In contrast, the goal of the mutation sampler is to describe not only what causal inferences people draw, but also how they do it.

5.2.3 | Mental Models Theory

The mutation sampler also shares some similarities with the *mental models theory* (MMT) of reasoning (Johnson-Laird, 1980). Both models posit that the fundamental units of reasoning are concrete possibilities and give special status to certain states (which in MMT are referred to as *initial mental models*). A more recent instantiation of MMT even defines a sampling process in which possibilities (i.e., mental models) are stochastically generated (Johnson-Laird et al., 2015; Khemlani et al., 2014). On this account, reasoners sample (with probability ϵ) from either the set of initial models or the set of *fully explicit* models (i.e., all models that are logically consistent with the premises). Model sampling continues until the number of models matches a number drawn from a Poisson distribution. Causal inferences are then drawn on the basis of the sampled models.

Despite these superficial similarities however, the mutation sampler and MMT differ in a number of important ways. One is that they posit different initial states.[14] A more fundamental difference is that states in MMT are qualitative possibilities consistent with the causal claims and as such do not have probabilities associated with them. Although probabilities can be derived from the sampled models (see Johnson-Laird et al., 2015), the manner in which those probabilities are computed treats the possibilities as equiprobable. For example, for the causal claim "insulting Adam causes Bethany to be angry", MMT stipulates that the fully explicit models are

- Adam was insulted and Bethany was angry
- Adam was not insulted and Bethany was angry
- Adam was not insulted and Bethany was not angry

This set of models implies that Bethany is angry two-thirds of the time, that the chance that Bethany is angry conditioned on Adam not being insulted is one half, and so forth. In contrast, we believe that when making such judgments human reasoners naturally take into account, for example, their beliefs about how often Bethany tends to be angry and the strength of the causal linkage between the insulting of Adam and Bethany's resulting anger. This intuition is reflected by the mutation sampler's fundamentally probabilistic representations that reflect variables' base rates, the strengths of the causal relations, and so forth. These factors influence causal judgments via the computation of the relative probability of states that occurs during the computation of Metropolis-Hastings transition probabilities.

These differences result in the models making different predictions for a number of key causal judgments. For one, MMT predicts no explaining away (Ali et al., 2011), a foundational signature of causal reasoning (Rottman & Hastie, 2016). In contrast, the mutation sampler can predict explaining away judgments, although weaker than those stipulated

[14] The initial states specified by the mutation sampler — the prototype states in which variables are either all present or all absent — do not generally correspond to either the initial or fully explicit mental models specified by MMT. For example, for the common cause network $Y_A \leftarrow X \rightarrow Y_B$ the initial models are $\{y_A^1 x^1 y_B^1\}$, that is, the single state in which all variables are present. The set of fully explicit models is formed by adding the logically possible states in which X is absent, resulting in $\{y_A^1 x^1 y_B^1, y_A^0 x^0 y_B^0, y_A^1 x^0 y_B^0, y_A^0 x^0 y_B^1, y_A^1 x^0 y_B^1\}$. For the common effect network $Y_A \rightarrow X \leftarrow Y_B$ the initial models are $\{y_A^1 x^1 y_B^1, y_A^1 x^1, x^1 y_B^1\}$ (in initial models only true states are represented, e.g. in model $y_A^1 x^1$ the state of Y_B is omitted rather than explicitly represented as false). The set of fully explicit models is formed by instantiating omitted states and adding the states in which Y_A and Y_B are absent, resulting in $\{y_A^1 x^1 y_B^1, y_A^1 x^1 y_B^0, y_A^0 x^1 y_B^1, y_A^0 x^0 y_B^0, y_A^0 x^1 y_B^0\}$. See Ali, Chater, and Oaksford (2011) for additional discussion. Note that the fact that the initial models for common cause and common effect networks differ illustrates MMT's assumption that even casual reasoners (those that reason on the basis of initial models alone) are sensitive to the direction of the causal links. In contrast, the mutation sampler assumes that initial states are independent of network topology.

by the normative model. MMT agrees with the normative model, and disagrees with the mutation sampler, by positing the absence of screening off errors in a common cause structure. As demonstrated in the "Empirical Tests" section, there is overwhelming evidence that people do in fact exhibit Markov violations in common cause networks.

Another important difference between models of course is that the mutation sampler but not MMT is an example of a rational process that approximates the normative standard to the extent that sufficient cognitive resources are available.

5.3 | New Empirical Questions

The primary focus of most studies of causal reasoning, including this one, has been on what causal inferences people draw and, in some instances, whether those inferences should be considered normative. By additionally asking *how* those inferences are drawn, we hope that the mutation sampler will expand the kinds of research questions that are posed and the kind of data that is considered. We now present a number of examples of these.

5.3.1 | Variability in Causal Judgments

One avenue of future research concerns the variability inherent in causal judgments. Although little past research has addressed this question, the recent study by Rottman and Hastie (2016) reported not only the mean responses to conditional probability queries but also histograms of those responses. For example, Figure 9A shows the histograms of responses to queries associated with a three variable common cause network like the one in Figure 4A from their Experiments 1A and 1B. [15] The first two columns are labeled "Markov" because they are queries relevant to evaluating independence violations, namely, $p(y_i^1 \mid x^1 y_j^0)$ and $p(y_i^1 \mid x^1 y_j^1)$. The third and fourth columns presents what Rottman and Hastie referred to as "middle" inferences (an inference to the "middle" X variable, i.e., $p(x^1 \mid y_A^1 y_B^1)$) and "transitive" inferences (i.e., $p(y_i^1 \mid y_j^1)$), respectively. In the figure, participants' ratings have been scaled into the range 0 to 1. A striking feature of Figure 9A is the large variability associated with each type of response. For example, the range that encompassed 95% of responses to the straightforward $p(y_i^1 \mid y_j^1)$ query was [.19, .96].

Figure 9B presents the corresponding histograms generated by the mutation sampler assuming a causal strength of *mean* (0.750, .875), a background strength of *mean* (.250, .125), and a chain length of 24. Note that whereas the predictions of the mutation sampler presented earlier in this paper were derived by computing the expected value of a joint distribution for a given chain length (and then the corresponding conditional probabilities), those in Figure 9B were computed by actually running the sampler 100,000 times. The figure shows that the predictions of the mutation sampler exhibits variability comparable to that of Rottman and Hastie's participants. It also tends to reproduce an unusual feature of their data, which is the presence of "spikes" in the histograms at .50. Such spikes, which might represent uncertainty in the mind of reasoners when reasoning about network states that have apparent inconsistencies (e.g., in $p(y_i^1 \mid x^1 y_j^0)$, the common cause X is present but its effect Y_j is absent) or include variables with unspecified values (e.g., in $p(y_i^1 \mid y_j^1)$ the state of X is unspecified). The mutation sampler predicts 0.50 for conditional probability queries when the networks states that are involved in the conditional probability calculation have not been visited by the stochastic sampling process and thus are at their initialized value. An interesting question for future research concerns whether Figure 9A reflects within or between participant variability. Assuming that each conditional probability judgment involves another run of the mutation sampler, it predicts both within and between participant variability.

[15] We thank Ben Rottman for providing the data from these studies. Experiments 1*A* ($N = 102$) and 1*B* ($N = 110$) were identical except that the strengths of the causal relations between X and the Ys (which were conveyed by participants observing individual cases) were 0.750 and .875, respectively, whereas the strength of alternative causes of the Ys were .250 and .125. We collapse across these similar studies in order to obtain a better estimate of the distribution

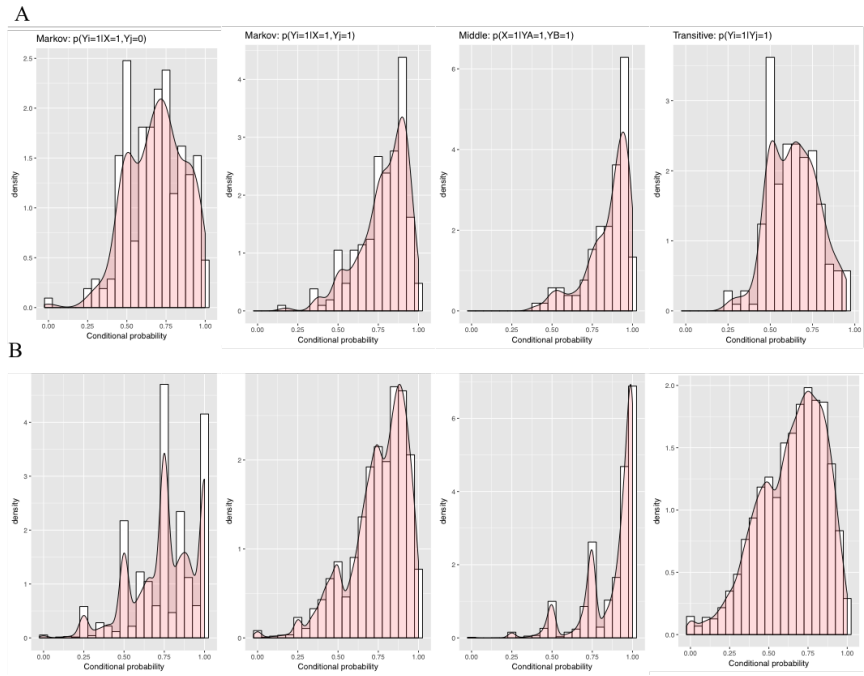


FIGURE 9 (A) Distribution of empirical ratings from Experiments 1A and 1B of Rottman and Hastie (2016). The first two columns present "Markov" inferences ($p(y_i^1 \mid x^1 y_j^0)$ and $p(y_i^1 \mid x^1 y_j^1)$), the third column presents "middle" inferences ($p(x^1 \mid y_A^1 y_B^1)$), and the fourth presents "transitive" inferences ($p(y_i^1 \mid y_j^1)$). The responses have in fact been collapsed over a number of different types of conditional probability queries that were considered equivalent for the purpose of the theoretical questions Rottman and Hastie addressed. For example, the ratings in the first column includes responses to both $p(y_i^1 \mid x^1 y_j^0)$ and $p(y_i^1 \mid x^0 y_j^1)$, where the latter rating was first flipped around the midpoint of the scale (0.50). Likewise, the second column includes $p(y_i^1 \mid x^1 y_j^1)$ and the flipped responses to $p(y_i^1 \mid X^0 y_j^0)$, the third includes $p(x^1 \mid y_A^1 y_B^1)$ and the flipped responses to $p(x^1 \mid y_A^0 y_B^0)$, and the fourth includes $p(y_i^1 \mid y_j^1)$ and the flipped responses to $p(y_i^1 \mid y_j^0)$. (B) The corresponding distributions generated by the mutation sampler.

5.3.2 | Individual Differences

There is also evidence for the presence of systematic differences in how individuals draw causal inferences. For example, Rehder (2014a) found that the causal inferences of a substantial minority of subjects exhibited no sensitivity to causal direction (treating, for example, common cause and common effect networks equivalently). These subjects, who Rehder dubbed *associative reasoners*, committed a large number of Markov violations.

Trueblood et al. (2017) also found differences in the magnitude of Markov violations and, moreover, that those violations correlated with other measures, such as the magnitude of order effects (i.e., the difference between $p(Z \mid X then Y)$ and $p(Z \mid Y then X)$) and measures they referred to as *reciprocity* and *memorylessness*. These measures in turn correlated with subjects' performance on the Cognitive Reflection Test (CRT), which is purported to measure differences in reasoners' tendency to emit intuitive versus deliberative responses (Frederick, 2005). That low CRT subjects committed more

of participants' responses (one based on $N = 212$).

causal reasoning errors was interpreted by Trueblood et al. as reflecting their tendency to adopt quantum probability representations with low dimensionality, which in fact tend to generate Markov violations, order effects, reciprocity, and memorylessness.

In the context of the mutation sampler, an obvious prediction is that individuals with a larger working memory capacity (and thus a better capacity to generate and maintain long sample chains) might exhibit fewer causal reasoning errors. Indeed, it has been shown that larger working memory capacity correlates with "analytical" thinking more generally (Evans & Over, 1996; Feeney, 2007; Stanovich & West, 1998; Stanovich, 1999). The fact that Markov violations correlate with the CRT (Trueblood et al., 2017), which in turn correlates with measures of general intelligence (Frederick, 2005; Toplak, West, & Stanovich, 2011), lends credence to this possibility. A related prediction would be larger causal reasoning errors for participants under working memory load.

5.3.3 | Reaction Times

Although reaction times have received relatively little attention in the causal reasoning literature, time pressure has been shown to increase people's tendency to emit intuitive versus deliberate responses in reasoning more generally (e.g., Evans & Curtis-Holmes, 2005; Evans et al., 2009; Finucane et al., 2000; Roberts & Newton, 2001). A prediction readily derivable from the mutation sampler is that time pressure will limit the number of samples and thus the accuracy of the inferences drawn. One study that addressed this question in a preliminary way is that of Rehder (2014a), who manipulated whether or not subjects were give a deadline to respond. In fact, Rehder found no impact of response deadlines on the magnitude of Markov violations. For a number of reasons however, this question deserves additional study. For example, the task in Rehder (2014a) was a relatively complex one in which subjects had to choose which of two scenarios was more likely to display a particular variable and this need to compare may have affected the mental processes invoked. It is also possible that the deadline used was not sufficient to induce the pressure needed to affect the subjects' sampling process.

5.4 | Generalizing to Additional Causal Scenarios

One strength of the current work is that it has tested the mutation sampler on a relatively large number of judgment types and causal network topologies. Yet, all those networks involved only generative causal relations between binary variables. Relatively little is known about the errors (e.g., independence violations) that reasoners commit with networks that include *inhibitory* causal relations. On one hand, applying mutation sampling to such networks is straightforward (one computes the Metropolis-Hastings ratio of the probability of two network states in the same manner as for networks with only generative relations). However, inhibitory relations raise the question of what constitutes the "prototype" states at which sampling should commence. For example, for the network $X \dashrightarrow Y \rightarrow Z$ (where $\dashrightarrow$ means that X inhibits Y), future empirical work might determine that reasonable prototype states are $x^1 y^0 z^0$ (the presence of X prevents Y and hence Z) and $x^0 y^1 z^1$ (the absence of X allows Y and hence Z).

The mutation sampler could also be applied to networks with continuous variables. Although most studies test binary variables, Rottman and Hastie (2016) found that the same pattern of independence violations obtains with continuous variables (e.g., in a $Y_A \leftarrow X \rightarrow Y_B$ common cause network, a larger value of Y_A led to a larger estimate of Y_B even when the state of X was known). Application of the mutation sampler to such cases would involve shifting probability mass in the joint probability density function; application of the sampling model would involve starting MCMC sampling at states in which X, Y_A, and Y_B all have either high values or low ones.

5.5 | Conclusion

Although the causal graphical model framework has served as a useful characterization of causal cognition, less work has investigated how such inferences can be drawn in a psychologically plausible manner. We have proposed that causal judgments are based on a relatively small number of samples drawn from a causal system and applied that approach to a large number of causal-based tasks, experimental conditions, and participants. We claim that the mutation sampler strikes an appropriate balance between the fact that people largely succeed at making sophisticated causal judgments while also committing systematic errors. And, we have argued that the computations and representations required by mutation sampling are within the reach of most reasoners for most types of judgments. We hope that the development of causal process models will expand the kinds of phenomena considered in the literature on causal cognition, including the variability of causal judgments, inter-individual differences in reasoning strategies, and reaction times.

6 | REFERENCES

Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, *119*(3), 403-418.

Anderson, J. R. (2013). *The adaptive character of thought*. Psychology Press.

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.

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference*. New York: Springer-Verlag.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from?. *Cognitive Psychology*, *96*, 1-25.

Evans, J. S. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.

Evans, J. S. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases believe bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, *11*, 382-389.

Evans, J. S. B. T., Handley, S. J., & Bacon, A. M. (2009). Reasoning under time pressure. *Experimental Psychology*, *56*, 77-83.

Feeney, A. (2007). Individual differences, dual processes, and induction. In A. Feeney & E. Heit (Eds.), *Inductive Reasoning* (pp. 302-327). Cambridge: Cambridge University Press.

Fernbach, P. M., & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument & Computation*, *4*, 64–88.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, *13*, 1-17.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*, 25-42.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357-364.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334-384.

Hagmayer, Y. (2016). Causal Bayes nets as psychological theories of causal reasoning: Evidence from psychological research. *Synthese*, *193*(4), 1107-1126.

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*, *3*(4), 254.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science*, *4*(1), 71-115.

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

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

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 461.

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

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856-876.

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.

Luhmann, C. C., & Ahn, W. K. (2007). BUCKLE: A model of unobserved cause learning. *Psychological Review*, *114*(3), 657.

Marr, D. (1982). *Vision.* San Francisco, CA: W. H. Freeman.

Mayrhofer, R., & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science*, *39*(1), 65-95.

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.

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology*, *67*(4), 186-216.

Park, J., & Sloman, S. A. (2014). Causal explanation in the face of contradiction. *Memory & Cognition*, *42*(5), 806-820.

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

Perales, J. C., Catena, A., & Maldonado, A. (2004). Inferring non-observed correlations from causal scenarios: The role of causal knowledge. *Learning and Motivation*, *35*(2), 115-135.

Rehder, B. (2003a). Categorization as causal reasoning. *Cognitive Science, 27*, 709-748.

Rehder, B. (2003b). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1141-1159

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

Rehder, B. (2014b). The role of functional form in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 670-692.

Rehder, B. (2018). Beyond Markov: Accounting for independence violations in causal reasoning. *Cognitive Psychology, 103*, 42-84.

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. & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General, 130*, 323-360.

Rehder, B. & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 659-683.

Rehder, B. & Kim, S. (2008). The role of coherence in causal-based categorization. In V. Sloutsky, B. Love, K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 285-290). Mahwah, NJ: Erlbaum.

Rehder, B. & Kim, S. (2010). Causal status and coherence in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1171-1206.

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.

Roberts, M. J., & Newton, E. J. (2001). Inspection times, the change task, and the rapid-response selection task. *Quarterly Journal of Experimental Psychology, 54*, 1031-1048.

Rottman, B., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140*, 109-139.

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.

Stanovich, K. E. (1999). *Who is rational?: Studies of individual differences in reasoning*. Psychology Press.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*, 161-188.

Tokplak, M.E., West, R.F., & Stanovich, K.E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-baises tasks. *Memory $ Cognition*, *39*, 1275-1289.

Trueblood, J. S., & Busemeyer, J. R. (2012). A quantum probability model of causal reasoning. *Frontiers in Psychology*, *3*, 1-13.

Trueblood, J.S., Yearsley, J.M., & Pothos, E.M. (2017). A quantum probability framework for human probabilistic inference. *Journal of Experimental Psychology: General*, *146*, 1307-1341.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*, *25*(1), 143-154.

Von Sydow, M., Hagmayer, Y., Meder, B., & Waldman, M. R. (2010, January). How causal reasoning can bias empirical evidence. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32, No. 32).

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.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645-647.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *31*, 216-227.

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.

Walsh, C. R., & Sloman, S. A. (2004, January). Revising causal beliefs. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 26, No. 26).

TABLE 2 Example Conditional Probability Distributions for the network in Figure 2.

Mutated variable	State of parent variables	Conditional probability expression	Conditional probability
a^1		c_A	.600
b^1		c_B	.600
c^1	$a^1 b^1$	$1 - (1 - b_C)(1 - m_{AC})(1 - m_{BC})$	.911
	$a^0 b^1$	$1 - (1 - b_C)(1 - m_{BC})$	.733
	$a^1 b^0$	$1 - (1 - b_C)(1 - m_{AC})$	.733
	$a^0 b^0$	b_C	.200
d^1	b^1	$1 - (1 - b_D)(1 - m_{BC})$	.733
	b^0	b_D	.200
e^1	c^1	$1 - (1 - b_E)(1 - m_{CE})$	.733
	c^0	b_E	.200

7 | APPENDIX A: COMPUTING METROPOLIS-HASTINGS TRANSITION PROBABILITIES

We present quantitative examples of Metropolis-Hastings ratios for the network in Figure 2. Doing so requires assumptions regarding the functions that link causes to their effects. One possibility, which corresponds to the conditions in the large majority of empirical studies that are considered later in this article, are that the causal links are *generative* (a cause makes its effects more likely) and *independent* (each causal link operates autonomously) and that multiple causal influences integrate according to a *noisy-or* function (Cheng, 1997). The generating function that defines the probability of variable V_i being present is then

$$\pi(v_i^1 \,|\, Pa(V_i)) = 1 - (1 - b_i) \sum_{V_j \in Pa(V_i)} (1 - m_{ji})^{ind(V_j)}$$

where $Pa(V_i)$ denotes the parents of V_i, m_{ji} is the strength of the causal link between parent variable V_j and V_i, $ind(V_j)$ is an indicator function that yields 1 if V_j is present and 0 otherwise, and b_i is the effect of background causes on V_i (causal influences exogenous to the model). If variable V_i is a "root" variable (i.e, has no causal parents), the probability that it is present is given by parameter c_{V_i}. Note that the principle of *causal sufficiency* that accompanies causal graphical models — that graph variables have no hidden causes in common — entails that root nodes are independent of one another.

Table 2 presents the probabilities of each variable in Figure 2 as a function of the state of its parents assuming that the marginal probabilities of the root causes A and B (c_A and c_B) are .60, that all causal relations (m_{AC}, m_{BC}, m_{BD}, and m_{AC}) have a strength of .67, and that the strength of the background causes for the remaining variables (b_C, b_D, and b_E) is .20. The probabilities in Table 2 constitute what is referred to as the *conditional probability distributions* (CPDs) for the network in Figure 2 under the given parameterization.

CPDs may be derived from generating functions other than a noisy-or of course. For example, rather than serving as independent causes of C in Figure 2, A and B might be *conjunctive causes* such that both need to be present in order to

generate C.

$$\pi(c^1|ab) = 1 - (1 - b_C)(1 - m_{ABC})^{ind(A)ind(B)}$$

where m_{ABC} is the strength of the conjunctive causal relationship relating A, B, and C. One empirical study considered in Appendix D instructed subjects on conjunctive causal relationships.

A causal graph's CPDs are sufficient to compute any marginal, conditional, and joint probability associated with that graph. For example, the Markov condition associated with causal graphical models stipulates that the graph in Figure 2 factors such that $p(abcde) = p(e|c)p(d|b)p(c|ab)p(a)p(b)$. Of course, any marginal and conditional probability can be derived from the joint.

We now show how a causal graph's CPDs are also sufficient to efficiently compute Metropolis-Hastings (MH) transition probabilities. Consider the first example MH ratio in Figure 2, $\pi(ce')/\pi(ce)$. Suppose that c^1e^0 holds in the current state q and the value of E has mutated to 1 in the proposed state q'. Noting that $\pi(ce) = \pi(e|c)\pi(c)$, the MH ratio can be computed from the CPDs in Table 2.

$$\frac{\pi(c^1e^1)}{\pi(c^1e^0)} = \frac{\pi(e^1|c^1)\pi(c^1)}{\pi(e^0|c^1)\pi(c^1)} = \frac{\pi(e^1|c^1)}{\pi(e^0|c^1)} = \frac{.733}{1 - .733} = 2.75$$

Table 3 presents the MH ratios for the cases in which E mutates to 1 as a function of the two possible values of C (c^0 and c^1).

The second example MH ratio in Figure 2 is $\pi(a'bc)/\pi(abc)$. Suppose that $a^0b^1c^0$ holds in the current state q and the value of A has mutated to 1 in the proposed state q'. Noting that $\pi(abc) = \pi(c|ab)\pi(a)\pi(b)$ and substituting in the appropriate CPDs from Table 2,

$$\frac{\pi(a^1b^1c^0)}{\pi(a^0b^1c^0)} = \frac{\pi(c^0|a^1b^1)\pi(a^1)\pi(b^1)}{\pi(c^0|a^0b^1)\pi(a^0)\pi(b^1)} = \frac{\pi(c^0|a^1b^1)\pi(a^1)}{\pi(c^0|a^0b^1)\pi(a^0)} = \frac{(1 - .911)(.60)}{(1 - .733)(1 - .60)} = 0.500$$

Table 3 presents the MH ratios for the case in which A mutates to 1 as a function of the four possible values of B and C.

The third example MH ratio in Figure 2 is $\pi(abc'e)/\pi(abce)$. Suppose that $a^0b^1c^0e^0$ holds in the current state q and the value of C has mutated to 1 in the proposed state q'. Noting that $\pi(abce) = \pi(e|c)\pi(c|ab)\pi(a)\pi(b)$ and substituting in the appropriate CPDs from Table 2,

$$\frac{\pi(a^0b^1c^1e^0)}{\pi(a^0b^1c^0e^0)} = \frac{\pi(e^0|c^1)\pi(c^1|a^0b^1)\pi(a^0)\pi(b^1)}{\pi(e^0|c^0)\pi(c^0|a^0b^1)\pi(a^0)\pi(b^1)} = \frac{\pi(e^0|c^1)\pi(c^1|a^0b^1)}{\pi(e^0|c^0)\pi(c^0|a^0b^1)} = \frac{(1 - .733)(.733)}{(1 - .20)(1 - .733)} = 0.917$$

Table 3 presents the MH ratios for the case in which C mutates to 1 as a function of the eight possible values of A, B, and E.

These examples illustrate how MH transition probabilities can be computed from a graph's CPDs. Notably, they can be computed without the full joint probability distribution.

TABLE 3 Examples of Metropolis-Hastings Ratios.

	State of Markov blanket	MH Ratio
$\pi(ce^1)/\pi(ce^0)$	c^1	2.75
	c^0	0.25
$\pi(a^1bc)/\pi(a^0bc)$	b^1c^1	1.86
	b^0c^1	5.50
	b^1c^0	0.50
	b^0c^0	0.50
$\pi(ac^1be)/\pi(ac^0be)$	$a^1b^1e^1$	37.58
	$a^1b^1e^0$	3.42
	$a^1b^0e^1$	10.08
	$a^0b^1e^1$	10.08
	$a^1b^0e^0$	0.92
	$a^0b^1e^0$	0.92
	$a^0b^0e^1$	0.92
	$a^0b^0e^0$	0.08

8 | APPENDIX B: CONVERGENCE PROPERTIES

This appendix justifies prototype states as a reasonable starting point given initial uncertainty about which states are highly probable for any particular graph. One consequence of generative causal relationships is that they yield positive correlations between variables. Thus, a reasonable first approximation of a graph's joint distribution is one in which graph states are more probable to the extent they exhibit *coherence*, variables that exhibit the same value. According to this approximation, prototype states are maximally coherent and thus serve as the best states at which to commence sampling. In fact, we show that this strategy often results in faster convergence to the normative distribution as compared to unbiased (uniformly randomly chosen) starting points. Of course, because cognitive resource limitations preclude large number of samples, this result implies that biasing the starting point results in more accurate causal inferences on average. Therefore, although the term "bias" may imply a maladaptive process, in fact it is consistent with our claim that people are capable causal reasoners that commit small but systematic errors.

We randomly generated 360,000 directed acyclic graphs of five binary variables. The graphs were generated by varying two factors—the strength and density of the causal relations—over 30 evenly spaced values from 0 to 1. A density value determined the number of causal links that were randomly chosen from the 10 causal links that define a five-variable network that is fully connected and includes no cycles. For example, the graph in Figure 2 has four causal relations and thus has a density of .40. Each cell in the 30 x 30 grid was sampled 400 times. In each graph, the marginal probability of a root cause (a variable that has no parent) was set to 0.5 and the strength of alternative causes of for each non-root variable was set to 0.1. Together these factors define a normative joint distribution over the five variables for each of the 360,000 graphs.

For each graph, two variants of the mutation sampler (biased or unbiased initialization) were then run with a chain length of exactly eight. For each of these (sampled) distributions, the KL divergence between it and the normative

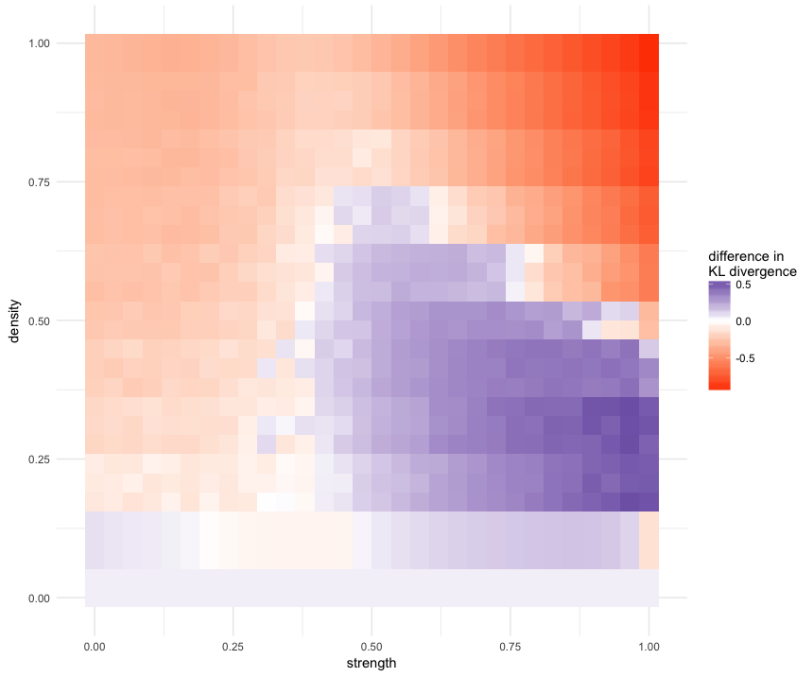


FIGURE 10 Heatmap of $\sqrt{D_{KL}(P||Q_{bias}) - D_{KL}(P||Q_{unbias})}$, where Q_{bias} refers to the estimated joint for a biased initialization. Red squares correspond to a benefit for biased initialization, and blue squares correspond to a benefit for unbiased initialization. Saturation corresponds to the magnitude of difference in KL divergence (larger differences have higher saturation).

distribution was computed. Figure 10 shows the square rooted difference in KL-divergence between the biased and unbiased initialization for each parameter combination.[16] Approximately 55% of the 360,000 runs resulted in faster convergence (i.e. lower KL-divergence) for the biased initialization.

The figure shows that biased initialization converged faster in dense graphs with strong causal links, because in such graphs the two prototypes will in fact be highly probable. In contrast, unbiased initialization converged faster in sparse graphs with strong causal links because the sparsity of the causal links means that maximally coherent items need not be the most probable. This result suggests that a biased initialization can be an effective approach given initial uncertainty about which graph states are most probable.

9 | APPENDIX C: ANALYSIS OF THE MUTATION SAMPLER'S PROPOSAL DISTRIBUTION

Rehder (2018) tested an extended common cause ($Z_A \leftarrow Y_A \leftarrow X \rightarrow Y_B \rightarrow Z_B$) and common effect graphs ($Z_A \rightarrow Y_A \rightarrow X \leftarrow Y_B \leftarrow Z_B$), shown in Figures 11A and 11B, respectively. These graphs provide an opportunity to test the mutation sampler's proposal distribution that specifies that only one variable can be mutated to generate a new proposal. To demonstrate

[16] We take the square root for visualization purposes only.

the importance of this proposal distribution, we defined an alternative sampler, which we name the *egalitarian sampler*, with a proposal distribution in which all network states have equal probability of being selected as a proposal, including those that differ from the current state on multiple variables.

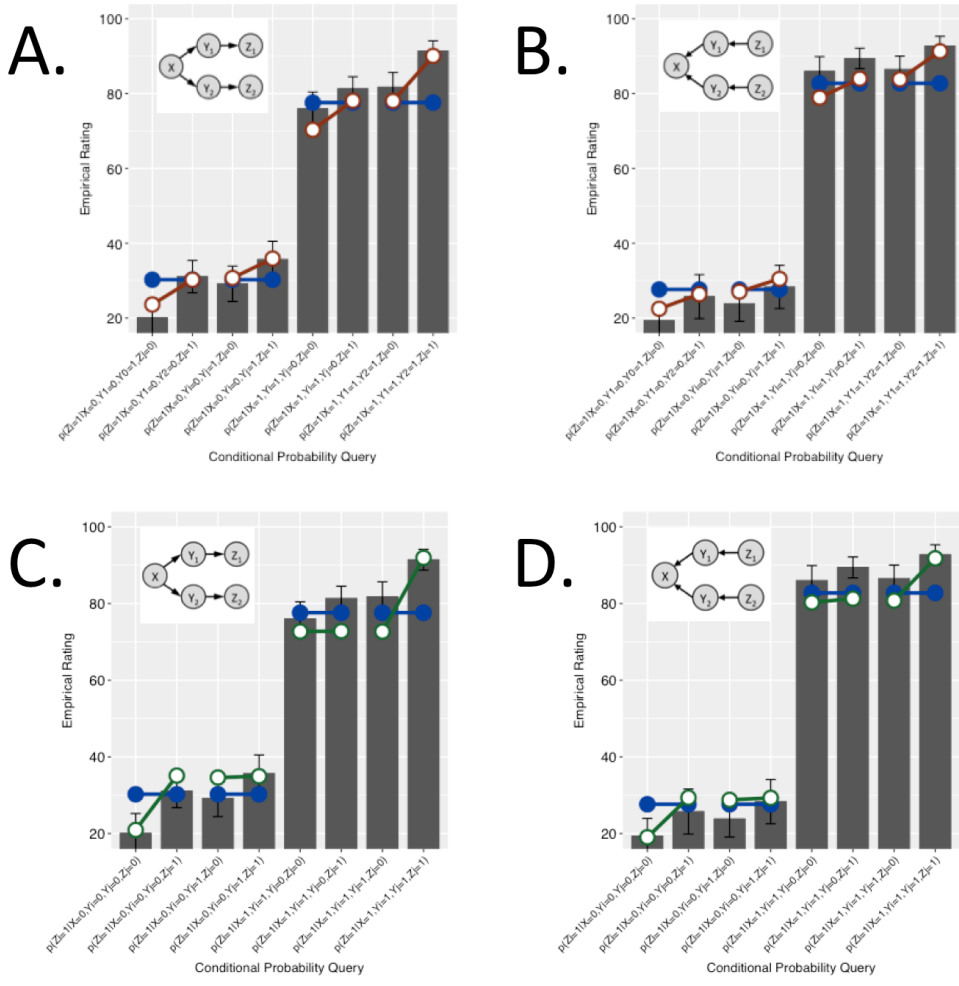


FIGURE 11 Evaluation of two alternative proposal distributions based on data from Rehder (2018, Experiment 2). Panels A and B present the fits of the normative model (blue lines, solid points) and mutation sampler to the extended common cause and common effect conditions, respectively. Panels C and D do the same but with the egalitarian sampler (green lines, open points).

We fit the normative model, the standard mutation sampler, and the new egalitarian sampler to the results from the extended common cause and common effect conditions from Experiment 2 of Rehder (2018) using the methods described in Appendix D. Figures 11A and 11B present the fits of the normative model and the mutation sampler superimposed on the empirical ratings from the extended common cause and common effect conditions, respectively. Figures 11C and 11D do the same with the egalitarian sampler.

The fits of the normative model in Figure 11 indicate that although each pair of adjacent conditional probability judgments should reflect independence (e.g., $p(z_i^1 | x^0 y_i^0 y_j^0 z_i^0) = p(z_i^1 | x^0 y_i^0 y_j^0 z_i^1)$), the subjects in fact violated independence. But while the fits of the egalitarian sampling model in Figures 11C and D reveal that it can account for some of those independence violations ($p(z_i^1 | x^0 y_i^0 y_j^0 z_i^0) < p(z_i^1 | x^0 y_i^0 y_j^0 z_i^1)$ and $p(z_i^1 | x^1 y_i^1 y_j^1 z_i^0) < p(z_i^1 | x^1 y_i^1 y_j^1 z_i^1)$), it cannot account for others ($p(z_i^1 | x^0 y_i^0 y_j^1 z_i^0) < p(z_i^1 | x^0 y_i^0 y_j^1 z_i^1)$ and $p(z_i^1 | x^1 y_i^1 y_j^0 z_i^0) < p(z_i^1 | x^1 y_i^1 y_j^0 z_i^1)$). In contrast, Figures 11A and 11B reveal that the mutation sampler provides a full account of the observed pattern of independence violations in this experiment.

Why do the mutation and egalitarian sampler differ in the independence violations they account for? They do so because they differ in their preferences for coherence in network states (where coherence refers to the proportion of variables in a state that have the same value). The mutation sampler instantiates a graded preference for coherence—the more variables with equal values, the better. This is so because, by making small adjustments to one of the (maximally coherent) prototype states, the mutation sampler also oversamples the mostly coherent neighboring states (those with 4 equal values) relative to other, less coherent states (those with 3 equal values). In contrast, because the egalitarian sampler can transition to any state, it does not oversample the mostly coherent states. As a result, although both models reproduce the independence violations in Figures 11 when the conditional probability queries involve a comparison between graph states with 4 and 5 equal values, only the mutation sampler does so when they involve states with 3 and 4 equal values. These findings provide empirical evidence in favor of the mutation sampler's proposal distribution.

10 | APPENDIX D: 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. Each of those conditions is described briefly in Figure 12 and in greater detail below.

Study	Expt.	Condition	No. of network variables	No. of subjects	No. of judgment types	Model	Parameters					Measures of fit		
							c	*m*	*b*	*λ*	*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%***
R18	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); R18 = Rehder (2018). 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 12 Comparison of normative model and mutation sampler for 4 existing causal reasoning datasets.

Rehder and Burnett (2005) (referred to as RB05 in Figure 12) taught participants categories whose features were causally related and then asked them 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 participant 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) (R14b in Figure 12) 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-participant 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 participant condition 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) (RW16) instructed participants 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-participant conditions in which participants 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 participants were only given a causal network.

Using the same materials as Rehder and Waldmann (2016), Rehder (2018) (R18) 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 of conditional probability judgments was 19 in Experiment 1 and 27 in Experiments 2 and 3. In Experiment 3 only participants were told that the causal relations operated 75% of the time.

In each condition participants' ratings were fit according to

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

where M is a model, θ_M are its parameters, t_i is a test item (i.e., a conditional probability query). Both the normative model and the mutation sampler were used to derive a joint probability distribution via the methods described in Appendix A, which was then used to derive the conditional probability appropriate for each test item. Although in general a causal graph's c parameters (representing the marginal probabilities of the root variables), m parameters (representing the strengths of the causal links), and b parameters (representing the strengths of causes extrinsic to the graph) can all differ from one another, the materials and counterbalancing used in the above studies were such that these parameters could be collapsed into a single c, a single m, and a single b (which, because they represent probabilities, were each constrained to the range [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 λ parameter representing the mean chain length, constrained to the range [2, 64]. The models were fit to each participant's causal judgments by identifying parameters that minimized squared error.

For each experimental condition, Figure 12 presents the number of variables in the network, the number of participants 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 participants 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[17], each averaged over the participants in that condition. Finally, the last column of Figure 12 presents the percentage of participants best fit by the mutation sampler in each condition.

[17] *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.

11 | APPENDIX E: MODEL FITS TO CAUSAL CATEGORIZATION STUDIES

The normative model and the mutation sampler were fit to causal categorization data from 25 experimental conditions reported in seven articles. In every condition, subjects were first taught categories whose binary features were causally related. The example of Myastars presented in the main text was one member of a set of six experimental categories that included biological kinds (species of ants and shrimp), non-living natural kinds (types of stars and molecules), and artifacts (types of cars and computers). To ensure that subjects learned the category's features and causal relations they were required to pass an extensive multiple choice test. The items presented on the subsequent classification test consisted of a set of features (e.g., a star with ionized helium, normal temperature, a large number of planets, etc.).

Study	Expt.	Condition	No. of network variables	No. of subjects	No. of judgment types	Model	Parameters							Measures of fit	
							c	m	m_2	b	λ	*bias*	γ	R	AIC
RH01	2	CC	4	78	8	Normative	0.940	0.756		0.444			0.166	0.988	25.4
						Mutation sampler	0.978	0.739		0.587	43.7	0.392	0.168	0.999	***11.9***
	2	CE-Indep.	4	78	8	Normative	0.821	0.615		0.317			0.166	0.998	12.3
						Mutation sampler	0.814	0.602		0.322	64.0	0.985	0.166	0.998	16.3
R03a		CC	4	36	8	Normative	0.895	0.733		0.294			0.170	0.981	28.1
						Mutation sampler	0.933	0.630		0.410	10.5	0.685	0.174	1.000	***-14.4***
		CE-Indep.	4	36	8	Normative	0.700	0.890		0.003			0.142	0.965	37.1
						Mutation sampler	0.623	0.833		0.004	4.2	0.902	0.143	0.997	***21.4***
R03b	1	Chain	4	36	16	Normative	0.931	0.613		0.735			0.164	0.986	41.6
						Mutation sampler	0.941	0.559		0.775	64.0	0.942	0.159	0.988	43.0
	2	Chain	4	36	16	Normative	0.722	0.867		0.258			0.255	0.910	71.7
						Mutation sampler	0.546	0.603		0.284	3.4	0.771	0.241	0.979	***53.2***
RK06	1	212	5	72	18	Normative	0.889	0.470		0.800			0.135	0.993	33.2
						Mutation sampler	0.584	0.308		0.424	3.0	1.000	0.129	0.998	***13.4***
		Chain	3	72	8	Normative	0.747	0.346		0.648			0.304	0.998	7.9
						Mutation sampler	0.772	0.271		0.698	64.0	0.210	0.304	1.000	***-2.8***
	2	311	5	72	16	Normative	0.862	0.448		0.830			0.144	0.987	42.9
						Mutation sampler	0.860	0.395		0.851	64.0	0.995	0.142	0.988	45.8
		Chain	3	72	8	Normative	0.718	0.331		0.602			0.333	0.982	22.4
						Mutation sampler	0.517	0.078		0.431	3.4	0.935	0.333	0.991	***20.8***
	3	113	5	72	16	Normative	0.925	0.473		0.686			0.171	0.964	56.5
						Mutation sampler	0.793	0.347		0.370	3.8	0.998	0.152	0.992	***36.8***
		Chain	3	72	8	Normative	0.799	0.522		0.561			0.318	0.985	24.4
						Mutation sampler	0.627	0.206		0.449	3.4	0.917	0.316	1.000	***-4.5***
RK08		Bipolar	4	36	16	Normative	0.871	0.920		0.320			0.269	0.908	77.1
						Mutation sampler	0.549	0.603		0.226	2.7	0.892	0.256	0.982	***56.2***
		Unipolar	4	36	16	Normative	0.939	0.545		0.709			0.180	0.972	52.6
						Mutation sampler	0.839	0.312		0.500	3.7	0.993	0.172	0.987	***44.8***
RK10	1	Chain (weak)	3	36	8	Normative	0.866	0.894		0.307			0.284	0.950	40.0
						Mutation sampler	0.647	0.555		0.310	2.0	0.882	0.271	0.997	***22.1***
		Chain (strong)	3	36	8	Normative	0.855	0.980		0.122			0.365	0.975	38.5
						Mutation sampler	0.623	0.816		0.255	2.0	0.859	0.372	0.994	***30.6***
	2	Chain (weak alt.)	3	36	8	Normative	0.834	0.899		0.187			0.352	0.974	35.7
						Mutation sampler	0.644	0.617		0.193	2.6	0.841	0.346	0.996	***25.4***
		Chain (strong alt.)	3	36	8	Normative	0.870	0.848		0.562			0.248	0.983	30.2
						Mutation sampler	0.587	0.665		0.332	2.7	0.965	0.228	1.000	***-15.1***
	App	Chain (no alt., weak)	3	36	8	Normative	0.915	0.936		0.156			0.337	0.979	35.9
						Mutation sampler	0.717	0.719		0.132	2.5	0.919	0.325	0.995	***28.2***
		Chain (no alt., strong)	3	36	8	Normative	0.802	0.999		0.005			0.315	0.989	35.3
						Mutation sampler	0.491	0.970		0.026	2.0	0.782	0.340	0.993	***35.3***
	App	Chain (60%)	3	48	8	Normative	0.888	0.891		0.374			0.290	0.963	37.9
						Mutation sampler	0.688	0.561		0.362	2.0	0.912	0.276	0.998	***16.6***
		Chain (90%)	3	48	8	Normative	0.865	0.758		0.432			0.316	0.962	36.1
						Mutation sampler	0.594	0.433		0.234	2.7	0.918	0.307	0.997	***19.8***
R14b	1	CE-Indep. & CE-Conj.	3	48	12	Normative	0.823	0.855	0.977	0.278			0.286	0.981	47.7
						Mutation sampler	0.729	0.862	0.970	0.150	29.1	1.000	0.267	0.992	***41.0***
	2	CE-Indep. (weak) & CE-Indep. (strong)	3	48	12	Normative	0.871	0.402	0.733	0.742			0.246	0.991	39.5
						Mutation sampler	0.861	0.382	0.721	0.735	64.0	1.000	0.245	0.990	44.4
		CE-Conj. (weak) & CE-Conj. (strong)	3	48	12	Normative	0.846	0.847	0.914	0.600			0.283	0.991	36.7
						Mutation sampler	0.853	0.819	0.899	0.632	47.1	0.863	0.281	0.994	***35.9***

Note. RH01 = Rehder & Hastie (2001); R03a = Rehder (2005a); R03b = Rehder (2005b); RK06 = Rehder & Kim (2006); RK08 = Rehder & Kim (2008); RK10 = Rehder & Kim (2010); 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 13 Fits of the normative model and the mutation sampler to past causal categorization studies. *AIC* values in bold italic red reflect cases in which the mutation sampler yielded a better fit as compared to the normative model.

Rehder and Hastie (2001, Experiment 2) (referred to as RH01 in Figure 13) taught subjects categories with four

binary features. In one condition the features formed a common cause network (one feature caused the other three). In another they formed a common effect network (one feature was caused by the other three). In all conditions subjects were told that one feature on each binary dimension occurred in 75% of category members and that the other occurred in 25% of category members (e.g., that 75% of Myastars have high density and 25% have normal density). Test items consisted of all 16 items that can be formed from four binary dimensions. For purposes of fitting, those 16 items were aggregated into 8 distinct types, formed by crossing the presence versus absence of X and the number of Y features present (0-3).

Rehder (2003a) (R03a) also taught subjects categories whose four features formed either a common cause or common effect network. Unlike Rehder and Hastie (2001), in this experiment subjects were not given the 75/25% feature base rate information. There were 16 test items of 8 distinct types. Rehder (2003b) (R03b) taught subjects categories whose four features formed a causal chain ($W \rightarrow X \rightarrow Y \rightarrow Z$). 75/25% feature base rate information was provided in Experiment 1 but not Experiment 2. There were 16 test items of 8 distinct types.

Rehder and Kim (2006) (RK06) taught subjects categories with five binary features. In Experiment 1, subjects performed two within-subject conditions that varied the causal knowledge that was provided. In one, the "212" condition, two features caused one feature, which in turn causes two features. In the other, three of the five features formed a causal chain (one of two sub-networks of the 212 network, counterbalanced over subjects, which ensured that comparisons between conditions involved the same category features. Experiment 2 consisted of a "311" condition (three features caused on feature, which caused another feature) and a causal chain (one of three sub-networks of the 311 structure). Experiment 3 consisted of a "113" condition (one feature causes one feature, which causes three features) and a causal chain (one of three sub-networks of the 113 structure). In all three experiments, one feature on each binary dimension was described as occurring in "most" category members while the other features was described as occurring in "some" category members. There were 32 test items, aggregated into various numbers of distinct types depending on the condition.

Rehder and Kim (2008) (RK08) taught subjects categories with four binary features that formed a causal chain. The unipolar condition was like the previous experiments in that the values on the binary dimensions were either "normal" or value that was distinct from normal (e.g., "Most Myastars have high density whereas some have normal density."). In contrast, in the bipolar condition there were two distinct (i.e., non-normal) values on each dimension (e.g., "Most Myastars have high density whereas some have low density."). There were 16 test items of 16 distinct types.

Rehder and Kim (2010) (RK10) taught subjects categories with three binary features that formed a causal chain ($X \rightarrow Y \rightarrow Z$). Experiment 1 included two conditions. In the strong condition, subjects were told that each cause feature brought about its effect 100% of the time. In the weak condition, they did so 75% of time. In the "weak-alt" condition of Experiment 2 subjects were told that each effect feature had no cause other than its parent in the causal chain. In the "strong-alt" condition they were told that each effect feature would occur with probability 50% even when its parent cause in the causal chain was absent. Two additional experiments testing a three-element causal chain were reported in Appendix B of Rehder and Kim (2010). The first manipulated causal strength between 100% and 75% (as in Experiment 1) but subjects were additionally told that each effect had no other causes. The second manipulated causal strength between 90% and 60%. In all experiments, one feature on each binary dimension was described as occurring in "most" category members and the dimensions were bipolar (as in Rehder and Kim, 2008, above). There were 8 test items, except for the final 90%/60% experiment in which there were 16 (each distinct test item was presented twice).

As described in Appendix D, Rehder (2014b) (R14b) taught subjects categories with six binary features. These features formed an independent and conjunctive common effect sub-network (Experiment 1), two independent common effect sub-networks (Experiment 2, independent condition), or two conjunctive common effect sub-networks (Experiment 2, conjunctive condition). Causal strength was manipulated within each condition of Experiment 2. Subjects

rendered not only conditional probability judgments (as described in Appendix D) but also categorization judgments. In each experiment there were 16 test items of 12 distinct types. The aggregate classification ratings in each condition were fit according to

$$rating(t_i) = 100 * p_{M,\theta_M}(t_i)^{\gamma}$$

where t_i is a test items, M is a model, and θ_M are its parameters. The normative model had the same causal model parameters as in Appendix D; the mutation sampler had an additional bias parameter (constrained to the range 0-1). Parameter γ (range 0-4) provides a nonlinear power transformation of a model's predicted joint probability[18]. The result was then multiplied by 100 to scale it onto the 0-100 rating scale. The model fitting procedure again minimized squared error. Figure 13, which has the same format as Figure 12 in Appendix D, presents the results of fitting the classification data.

12 | APPENDIX F: SAMPLING FOR CONDITIONAL PROBABILITY QUERIES

Figure 14 presents the rates of convergence for six conditional probability queries associated with the five-variable network presented earlier in Figure 2 as a function of chain length (λ). Each panel shows the conditional probability computed by the normative model (red), the mutation sampler (black), and the alternative sampler in which states in which the conditional probability query's antecedent are true are ten times more likely to be sampled than those in which it is false (blue). In each panel the alternative sampler converges faster than the standard one. Moreover, this effect is larger as the number of variables involved in the antecedent increases: The advantage for the alternative sampler is larger in the bottom row in Figure 14, which presents conditional probability queries whose antecedents include four variables, than it is in the top row, which presents queries whose antecedents include only two variables. These results that the mutation sampler can compute reasonably accurate conditional probability queries in a large network with even a modest number of samples.

[18] The power transformation allows the evidence for category membership provided by each feature in a test item to be integrated in a manner other than multiplication. Examination of Table C1 indicates that the fits yielded values of γ in the range [0.1, 0.4]. The transformation specified by a γ in this range is very similar to a logarithm. That is, subjects tended to add rather than multiply the evidence for category membership provided by each feature. See Griffiths and Tenenbaum (2005) and Rehder (2014b) for examples of power function transformations.

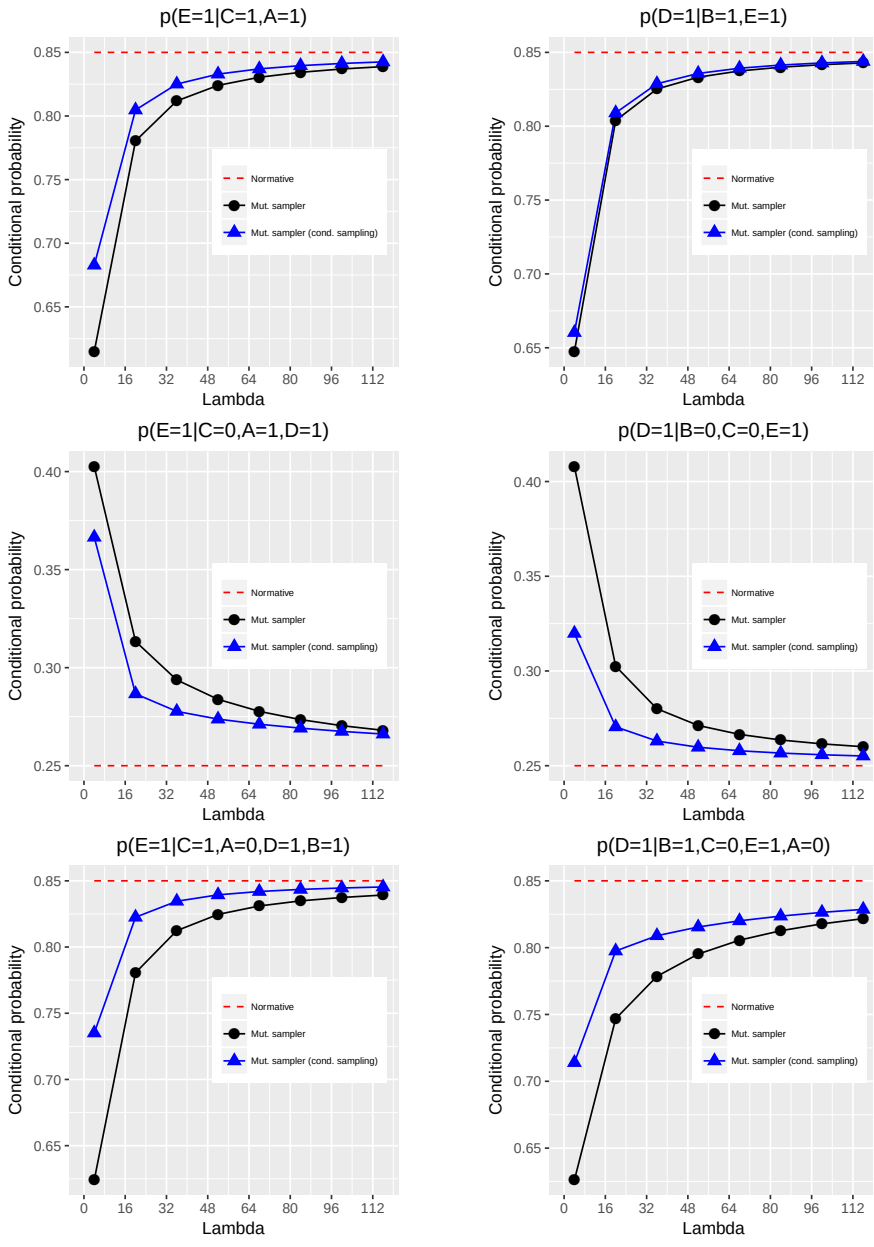


FIGURE 14 Six conditional probability judgments generated by the normative model (red), the standard mutation sampler (black), and the mutation sampler with an alternative proposal distribution (blue). Marginal probability of root causes = .50; causal strengths = .80, and strength of alternative causes = .25.