Cognitive Psychology 87 (2016) 88–134

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away [☆]

Benjamin M. Rottman [a,*], Reid Hastie [b]

[a] *Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, PA 15260, United States*
[b] *The University of Chicago Booth School of Business, 5807 South Woodlawn Ave, Chicago, IL 60637, United States*

ARTICLE INFO

Article history:
Accepted 7 May 2016
Available online 1 June 2016

Keywords:
Causal inference
Bayesian Networks
Markov Assumption
Explaining away
Experience

ABSTRACT

Making judgments by relying on beliefs about the causal relationships between events is a fundamental capacity of everyday cognition. In the last decade, Causal Bayesian Networks have been proposed as a framework for modeling causal reasoning. Two experiments were conducted to provide comprehensive data sets with which to evaluate a variety of different types of judgments in comparison to the standard Bayesian networks calculations. Participants were introduced to a fictional system of three events and observed a set of learning trials that instantiated the multivariate distribution relating the three variables. We tested inferences on chains $X_1 \rightarrow Y \rightarrow X_2$, common cause structures $X_1 \leftarrow Y \rightarrow X_2$, and common effect structures $X_1 \rightarrow Y \leftarrow X_2$, on binary and numerical variables, and with high and intermediate causal strengths. We tested transitive inferences, inferences when one variable is irrelevant because it is blocked by an intervening variable (Markov Assumption), inferences from two variables to a middle variable, and inferences about the presence of one cause when the alternative cause was known to have occurred (the normative "explaining away" pattern). Compared to the normative account, in general, when the judgments should change, they change in the normative direction. However, we also discuss a few persistent violations of the standard normative model. In addition, we evaluate the relative success of 12 theoretical explanations for these deviations.

[☆] *Author note:* This research was supported by NIH grant 1F32HL108711 and the University of Chicago Booth School of Business.

[*] Corresponding author.

http://dx.doi.org/10.1016/j.cogpsych.2016.05.002
0010-0285/© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Causal inference is a ubiquitous aspect of every-day life. How much will Apple Computer's stock increase if it releases a new model of iPhone before the holiday season? What if Samsung also releases a new model? What are my chances of developing Sickle Cell disease given that my mother has Sickle Cell disease? Does the probability increase if her mother also had Sickle Cell disease? How much better will I perform on the exam if I study one more hour? What if that means that I will get one hour less sleep? We make hundreds of judgments every day that rely on beliefs about how two or more events are causally and probabilistically related to each other.

Some theorists have even suggested that causal cognition is fundamental to almost all everyday and expert judgments (Hagmayer & Osman, 2012; Hastie, 2016). Many well-established phenomena from the literature on judgment and decision making are directly produced or are moderated by causal reasoning, including multiple-cue judgments, the reliance on base rate information in judgments under uncertainty, hindsight and belief perseveration, conjunction fallacies, the Planning Fallacy, and many aspects of consumer judgments. Many decisions can also be best understood as choosing actions because they are expected to cause desired outcomes (Hagmayer & Sloman, 2009). And, there are many empirical demonstrations that category classification and category-based inferences are saturated with causal reasoning (Murphy & Medin, 1985; Rehder, 2010).

The current research focuses on judgments that people make about multiple causal events embedded in a causal structure. For example, when predicting whether one has Sickle Cell disease from knowledge of one's mother's and grandmother's status, the following causal structure can be used to guide the prediction: [*Grandmother Has Sickle Cell Disease* → *Mother Has Sickle Cell Disease* → *Child's Sickle Cell Status*]. The judgment about whether to study for another hour for a test also makes use of causal structure knowledge; studying for an extra hour causes less sleep, and both the amount of studying and the amount of sleep influence exam performance. Many of the consequential judgments people make involve events embedded in networks.

The outline for the introduction is as follows. We first introduce the basic Causal Bayesian Network model of causal judgment. We then discuss limitations of prior experiments that have tested the CBN model. Finally, we discuss prior research on three reasoning habits that will be the focus of the current research.

1.1. The simple point-estimate Causal Bayesian Network (CBN) model

In the last two decades Graphical Probabilistic Models (also called Causal Bayesian Networks; CBN) have come to dominate the modeling of probabilistic causal phenomena in science, engineering, and medicine, and they have also become the most popular model of human causal reasoning (Holyoak & Cheng, 2011; Rips, 2008; Sloman & Lagnado, 2015; Waldmann & Hagmayer, 2013). Causal Bayesian Networks are specifically designed to handle inference problems for which the events are embedded within a causal network.

CBN theory works at both a qualitative and quantitative level. Qualitatively, the structure of the network can be used to deduce certain properties. For example, when predicting X_1 on the common cause network $[X_1 \leftarrow Y \rightarrow X_2]$, once the state of Y is known X_2 is completely irrelevant. Additionally, the correlation between X_1 and Y must be stronger than the correlation between X_1 and X_2. Some of the most influential studies of inferences on causal networks have focused on qualitative judgments (Park & Sloman, 2013, except Experiment 3; Rehder, 2014; Rehder & Burnett, 2005).

CBN theory also can be used to make quantitative inferences such as estimating precisely the probability of X_1 given knowledge that Y is present, summarized as $P(x_1 = 1|y = 1)$. Making these quantitative inferences requires knowledge of the parameters that define the statistical relations between each cause–effect link in the network. There are two ways that such parameters can be conveyed to participants. The first is to verbally state the probability of each effect given its direct causes, or equivalently, the 'causal strength' that each cause has on its direct effects (Fernbach, Darlow, & Sloman, 2010; Fernbach & Rehder, 2013; Krynski & Tenenbaum, 2007; Morris & Larrick, 1995). Participants then could deduce inferences like $P(x_1 = 1|y = 1)$ by mentally applying Bayes' rule and other principles of probability.

Another way to convey the parameters to participants is to have them experience a set of learning trials (Edgell, Harbison, Neace, Nahinsky, & Lajoie, 2004; Meder, Hagmayer, & Waldmann, 2009; Park & Sloman, 2013; von Sydow, Meder, & Hagmayer, 2009; Waldmann & Hagmayer, 2005). On each trial, each of the three variables can be present or absent, and the entire set of trials instantiates the multivariate distribution of the three variables. This multivariate distribution is 'faithful' to the structure in that it upholds the qualitative properties implied by the structure. When a learner experiences the multivariate distribution, there are two mathematically equivalent versions of this model. The learner could use the statistical relations among pairs of cause–effect nodes to infer causal strength parameters (e.g., Cheng, 1997), and then use the parameters to make predictions about other variables. Alternatively, the learner could make inferences directly from the learning data. For example, if inferring $P(x_1 = 1|y = 1)$, the learner could search the number of trials in which $x_1 = 1$ and $y = 1$, and divide this number by the number of trials in which $y = 1$ (see Rottman & Hastie, 2014 for a tutorial and discussion of these options).

There are several possible ways to elaborate the basic model, which we consider in Section 4. Our point is that even though alternative normative models could be entertained, there is a sizeable literature using the point-estimate model as the normative model.

1.2. Limitations of prior research

Over the past 15 years there have been a number of studies that have investigated how people make predictions for events that form a causal network, especially the chain $X_1 \rightarrow Y \rightarrow X_2$, common cause $X_1 \leftarrow Y \rightarrow X_2$, and common effect $X_1 \rightarrow Y \leftarrow X_2$. These studies have produced a solid foundation of empirical findings, and a useful basic methodological framework (see Rottman & Hastie, 2014 for a review). However, most of these studies have used similar methodologies, and the similarity means that they have a shared set of limitations. The present experiments consolidate, refine, and extend this important research program.

1.2.1. Comprehensiveness of research findings

Our first contribution is combine and extend methods and practices from many prior investigations into more comprehensive experimental designs. On any given causal structure such as the chain, $[X_1 \rightarrow Y \rightarrow X_2]$, there are a great variety of different inferences that an individual can make: a 'one-link' inferences from one event to an adjacent event such as $P(X_1|Y)$,[1] 'transitive' inferences $P(X_1|X_2)$, inferences about the middle event given the two flanking events $P(Y|X_1, X_2)$, and inferences to a terminal event given two other events $P(X_1|Y, X_2)$. Most prior studies have examined a subset of these inferences, and often on one or two causal structures. Searching for patterns of performance across inference types and structures has required comparisons across studies. We also believe that the inference of $P(X_1|Y, X_2)$, on the common effect structure $[X_1 \rightarrow Y \leftarrow X_2]$ has not received enough attention in carefully controlled studies (we discuss this inference more thoroughly below). In the current experiments we study all of these inferences, to have a more complete understanding of performance on different inferences in a single paradigm.

Another limitation is the almost exclusive focus in research, on binary, occur/does-not-occur events. However, many real-world inferences (e.g., 'how well will I perform on a test if I study for another hour') involve continuous or at least interval-level conceptions of cause and effect *magnitudes*. Another motivation for testing numerical variables is that several findings with binary variables can be explained by weak inferences when two cues contradict each other (e.g., when inferring X_1 on the chain $[X_1 \rightarrow Y \rightarrow X_2]$, if $y = 1$ but $x_2 = 0$), and overly strong inferences when two cues are consistent (e.g., if $y = 1$ and $x_2 = 1$). When the three variables are numerical (e.g., can take on values 1–100), these issues of consistent versus inconsistent cues do not play out the same way because the two cues will rarely have exactly the same value (see Section 1.3 for further discussion).

[1] For readers less familiar with probability notation, $P(X_1|Y)$ means the probability of X_1 given knowledge of the state of Y.

1.2.2. Finer-grained analysis of results

In evaluating whether the inferences correspond to the Causal Bayesian Network theory, most of the prior research has focused on qualitative comparisons, for example, testing whether the judgments for one inference are on average higher than the judgments for another inference. These qualitative comparisons provide insights into many of the implications of CBN theory. However, a finer-grained analysis has the potential to reveal additional habits of reasoning that are not revealed by qualitative comparisons. In the current research, we examined not only patterns in the average judgments, but also the distributions of those judgments. To presage the results, we found some consistent patterns of spikes and asymmetries, that support alternative theories that have previously not been proposed to account for judgments on causal networks.

In the current research we tested both qualitative comparisons, such as whether judgments for inference A are on average different from judgments for inference B, but also how close these judgments are to the quantitative predictions made by CBN theory. We acknowledge that some readers will only care about the qualitative predictions, and believe that CBN theory (like many other theories in cognitive psychology) should not be applied to make quantitative predictions. One principled justification for the focus on qualitative relationships is that there is abundant evidence that people do not conceptualize probabilities in a manner that maps linearly onto a true probability metric (e.g., the tendency to over-estimate low probabilities and under-estimate high probabilities, Gonzalez & Wu, 1999; Zhang & Maloney, 2012). Thus, any observed violations of quantitative predictions from CBN may derive from non-standard thinking about probabilities, not to something essential about causal reasoning.

Nevertheless, we continue to believe that a fine-grained analysis that compares the judgments to the quantitative predictions of CBN theory can be useful. In fact, we find that many judgments are too weak, too close to the middle of the probability scale. However, we argue that the underlying cause of this weakness is not simply a non-linear subjective probabilities representation, but rather that certain types of judgments are too weak and others are *too strong*. We also conclude that studying inferences on both qualitative, binary representations and quantitative numerical representations yields important insights into the deeper underlying reasoning habits.

Another justification for finer-grained analysis is that certain qualitative patterns can be explained both by CBN and also by simple judgment heuristics. For example, consider the structure $[X_1 \rightarrow Y \rightarrow X_2]$, and assume that both causal relations are positive. Because Y is closer to X_1 than X_2 is to X_1, one might infer that that if $x_1 = 1$, then y is probably also 1, but be less certain about the value of x_2. This inference is justifiable both by a simple proximity heuristic (Burnett, 2004) as well as CBN. Since it is possible to approximate the CBN predictions with some simple heuristics, we think it is important to examine the distributions of responses and the quantitative fit with predictions.

One of the reasons that the previous research has tended not to consider the fine-grained quantitative predictions of CBN theory is that many of the previous studies have given participants a causal network, but have not provided additional statistical information with which to make inferences. In some cases this meant that there was no normative numerical calculation to which human judgments could be compared, only qualitative response patterns could be assessed.

In the current experiments we provide participants with trial-by-trial learning experiences that support quantitative inferences. Aside from just allowing quantitative judgments, there is another benefit of learning experience: Since the learning data are faithful to the causal structure, experiencing the learning data should improve judgment performance even on qualitative assessments. This means that previous studies that did not include learning data may have underestimated human reasoning capacities.

1.2.3. Consideration of alternative explanations

Lastly, by implementing a comprehensive experimental design, and applying a finer-grained scrutiny to the judgments made by participants, the current research provides a broad empirical basis for the assessment of more theoretical accounts of causal reasoning. There have been a few sophisticated proposals of modifications to the CBN framework, with accounts for some deviations from the basic model (Park & Sloman, 2013; Rehder, 2014). Our research builds on those proposals by (1) collecting a new and extensive set of empirical results, and (2) proposing some new accounts informed by these results. In Section 4 we assess 12 different accounts of 5 key findings, to provide comprehensive comparative evaluations of all current theoretical proposals, as well as explorations of how combinations of those accounts may be sufficient to explain the results.

In summary, the broad goals of the current research are to build off of the important contributions from the previous research (1) by testing many of the same phenomena in more comprehensive experimental designs, (2) by applying more scrutiny when comparing the results to the CBN theory, and (3) by comparing the results to the large set of previous accounts, as well as proposing some additional plausible accounts.

1.3. Specific motivation for current experiments

In a recent review of previous studies of inferences from causal structures, we concluded that even though people follow many of the predictions of the CBN model, people tend to make three systematic and prevalent errors when their judgments are compared to the maximum likelihood point-estimate prescriptions of the CBN model (Rottman & Hastie, 2014). However, that paper was a review of earlier empirical studies but did not conduct original experiments. In the following three sections we summarize the evidence for these three reasoning habits, and note some limitations on past research, to further motivate the present experiments.

1.3.1. The Markov Assumption

When making inferences about one variable on a causal structure, certain other variables are irrelevant for making the inference. For example, simple autosomal recessive genetic diseases like Sickle Cell Disease have the property that if a mother has the disease, whether or not her parents have the disease is irrelevant for calculating the likelihood of her children having the disease. More generally, on the chain $[X_1 \rightarrow Y \rightarrow X_2]$, X_1 is irrelevant when inferring X_2 once the state of the mediator (Y) is known, yet people often behave as if X_1 is still relevant for inferring X_2 above and beyond Y.

More technically, the Markov Assumption states that for a particular variable (e.g. X_2) in a causal structure (e.g., $[X_1 \rightarrow Y \rightarrow X_2]$) once one conditions on the direct causes (Y) of the variable (X_2), all variables except for direct and indirect effects of X_2 are independent of X_2. In this example, X_2 is conditionally independent from X_1 once Y (the only cause of X_2) is conditioned upon. If the variables are binary, the Markov Assumption is that $P(x_2 = 1|y = 1,\ x_1 = 1) = P(x_2 = 1|y = 1,\ x_1 = 0)$, and that $P(x_2 = 1|y = 0, x_1 = 1) = P(x_2 = 1|y = 0, x_1 = 0)$.[2] Restated, the probability of X_2 is not influenced by the state of X_1 once the state of Y is known. Conditional independence of X_1 and X_2 given Y on a causal chain is the standard definition of a mediator. Conditional independence is symmetric, which means that X_1 is not influenced by the state of X_2 once the state of Y is known. This is the reason that X_1 and X_2 are both labeled as X even though one is downstream from the other. In the rest of the paper, when inferences are normatively symmetric for the two Xs, the subscripts i and j are used to represent the two Xs; the Markov Assumption can be rewritten as $P(x_i = 1|y = 1, x_j = 1) = P(x_i = 1|y = 1, x_j = 0)$. The Markov Assumption also works on the common cause structure $[X_1 \leftarrow Y \rightarrow X_2]$; X_1 and X_2 are conditionally independent once Y is known. For the common effect structure $[X_1 \rightarrow Y \leftarrow X_2]$, X_1 is unconditionally independent of X_2. The reason is that neither X_1 or X_2 have any known causes, so they must be independent of each other even without conditioning on any other variables.

The Markov Assumption is critical for the Causal Networks framework because it identifies which nodes are relevant or irrelevant when making a particular inference. For example when inferring X_2 on the structure $[X_1 \rightarrow Y \rightarrow X_2]$, X_1 is relevant if the state of Y is not known, but X_1 is irrelevant if the state of Y is known. The Markov Assumption becomes even more important with larger networks because the Markov Assumption may identify many of the nodes as irrelevant for a given inference, which can vastly simplify calculations (especially important in applications to complex, engineered systems).

Previous research on whether people use the Markov Assumption when making inferences on causal networks has mainly relied on a method of telling participants a cover-story about a causal

[2] As a reminder for readers less familiar with probability notation, the term $P(x_2 = 1|y = 1, x_1 = 1)$ can be read as, "the probability that $x_2 = 1$ given that $y = 1$ and $x_1 = 1$"; or alternately as, "the probability that x_2 occurs given that y occurs and x_1 occurs."

structure such as $[X_1 \rightarrow Y \rightarrow X_2]$ and asking them to make inferences like $P(x_i = 1|y = 1, x_j = 1)$ and $P(x_i = 1|y = 1, x_j = 0)$, which should be equivalent according to the Markov Assumption. But, the usual finding is that people infer higher probabilities for $P(x_i = 1|y = 1, x_j = 1)$ than $P(x_i = 1|y = 1, x_j = 0)$. One of the weaknesses of this method is that participants might import their own beliefs about the scenario, which could change the causal relations provided in the experimenter's cover-story. For example, believing that there is an additional direct causal relation $[X_1 \rightarrow X_2]$ (not mentioned by the experimenter) would justify the higher ratings for $P(x_i = 1|y = 1, x_j = 1)$. Researchers have used a variety of approaches such as basing the cover-story on unfamiliar variables (Rehder & Burnett, 2005) and clever counterbalancing (Rehder, 2014) to minimize the possibility that background knowledge could produce apparent Markov violations.

Another limitation with the cover-story methodology is that conveying complex probabilistic information to statistically-naïve participants is challenging. For example, does telling the participants that, "X_1 causes Y and Y causes X_2" imply that X_1 does not influence X_2 above and beyond the relationships with Y? (Anecdotally, when teaching students about causal structures and mediation, conditional independence is not always intuitively obvious from the structural diagram.)

Lastly, unless specific parameters of the causal structure are conveyed to participants, no precise normative answers for the inferences can be calculated; only qualitative patterns in the inferences can be inferred.

The Markov Assumption has rarely been tested when participants experience the probabilistic relationships between the variables over a sequence of sets of events based on the relationships (but see Park & Sloman, 2013, Experiment 3). This is surprising because a learning phase, in which participants experience the multivariate distribution of events, is a common procedure in studies of causal learning. Furthermore, experience is known to reduce other biases in probabilistic inference such as neglect of base rates in diagnostic inferences (Christensen-Szalanski & Beach, 1982).

It is possible that experience with a representative sample of events would reduce or eliminate violations of the Markov Assumption. Thus, in the present experiments we test whether our participants' inferences respect or violate the Markov Assumption after experiencing the multivariate distribution of events. In Experiment 1 we present scenarios with binary events, and in Experiment 2 we present scenarios with numerically-valued variables. Events defined on numerical variables (or subjective magnitudes that could be modeled as quantities) are common in everyday situations, yet they have rarely been studied in research on causal reasoning. Furthermore, numerical variables might be interpreted by participants to imply more precision in measurement compared to binary variables, which could lead to fewer Markov violations.

1.3.2. The strength of inferences

In our previous review (Rottman & Hastie, 2014) we found that many inferences on causal structures tended to be too weak. For example, Meder, Hagmayer, and Waldmann (2008, Experiment 1) told participants about four chemicals that might be present in wine $[A, B, C, D]$ and explained the causal relations between the chemicals $[B \leftarrow A \rightarrow C$ and $B \rightarrow D \leftarrow C]$. Participants then observed whether the chemicals were present or absent in each of 40 casks of wine and made inferences like $P(d = 1|c = 1)$ and $P(d = 1|c = 0)$. Converted to a probability scale, participants should have answered .90 and .13, but instead they answered, on average, .74 and .37. The inferences almost always moved in the normatively expected directions; however, they moved too little. Still, participants seemed to be paying close attention to the specific parameter values (also see Meder et al., 2009). For example, they were remarkably sensitive to variations in the base rates and the strengths of the various causal relations. A comprehensive review found similar patterns of weak inferences in many other studies, but there was wide variation in methods and statistical tests were rarely performed to test whether the inferences deviated from the normative calculations (Rottman & Hastie, 2014). And, there is prior work (e.g., Phillips & Edwards, 1966) finding conservatism in other Bayesian updating tasks with binary variables.

Evidence from related tasks involving inferences with numerical instead of binary variables is mixed. Kahneman and Tversky (1973) reported studies in which participants' statistical inferences were too strong. Participants were asked to estimate a student's GPA from three predictors with varying degrees of informativeness for GPA, where informativeness could be interpreted as causal relevance. For a strong predictor, a high score should predict a high GPA. But for a moderately informative predictor a high score should predict a slightly high GPA, a statistical pattern known as "regression" (Campbell & Kenny, 1999). Kahneman and Tversky found that the regressiveness of the three predictors was very small, in some cases non-existent. They explained this non-regressive habit with the "representativeness" heuristic; arguing that a high score on the predictor is most representative of a similarly high GPA.

In contrast, Lichtenstein, Earle, and Slovic (1975) trained their participants extensively on a linear prediction task, giving them practice with feedback, and found that people's inferences were overly regressive (conservative or weak in our terms). Yates and Jagacinski (1979) also gave participants trial-by-trial learning experience but no feedback and found that their predictions were regressive, though it was unclear whether they were normatively regressive.

In sum, it is not clear from the broader literature if prediction on numerical variables tends to be too strong or too weak. Furthermore, most of the previous work on regressiveness studied inferences on simple structures with one or two causes on an effect (e.g., inferences of Y on one-link $[X \to Y]$ and common effect $[X_1 \to Y \leftarrow X_2]$ structures), but not on chains $[X_1 \to Y \to X_2]$ and common cause $[X_1 \leftarrow Y \to X_2]$ structures. And, in those studies the role of causal reasoning in the judgments was implicit or perhaps not involved at all.

The current experiments were designed specifically to test whether inferences on three-node causal networks with a case-by-case learning experience are normative, too strong, or too weak. We examined transitive inferences such as from X_1 to X_2 on the chain $[X_1 \to Y \to X_2]$, and inferences from two variables to a "middle" variable such as inferring Y given X_1 and X_2. In Experiment 1 we test these patterns of inferences on causal scenarios defined on binary events and in Experiment 2 on numerical causal variables. Based on prior results, we predicted that inferences would be too weak with binary variables, but would be too strong for numerical variables because participants might anchor on a specific cue value and fail to adjust (regress) sufficiently towards the mean (as in Kahneman & Tversky's, 1973 studies).

1.3.3. "Explaining away" inferences on common effect $[X_1 \to Y \leftarrow X_2]$ structures

Explaining away has long been viewed as an underlying principle in situations in which there are multiple causes of an effect, and the goal is to figure out which of the causes is responsible for the effect such as in social attribution (Jones, 1979; Kelley, 1972), legal exoneration, and medical diagnosis.[3] For example, if we believe that Tim's poor score on the test could have been caused by the noisy construction outside the classroom (situation), then we are less likely to infer that he has a low aptitude (disposition). In contrast, if Tim scored poorly even when there was no construction, then we are more likely to infer that it is due to low aptitude. Similar cases arise in medical diagnosis. Upon encountering a patient with a cough, if we know that the patient has asthma, it is not necessary to infer that the patient also has the flu. However, if we know that the patient does not have asthma, we are more likely to infer that the patient has the flu. Both of these examples involve situations in which there are two causes of one effect. Pearl assumed explaining away to be such a "prevailing pattern of human reasoning" that he used it to motivate a normative-mathematical explanation (Pearl, 1988, p. 49).

Table 1 shows hypothetical data with two causes (X_1 and X_2) that are independent, $P(x_i = 1|x_j = 1) = P(x_i = 1|x_j = 0) = .50$, and both of which on their own produce the effect (Y) with a probability of .50. It is assumed that they combine through a "Noisy-Or" integration function so that the likelihood of the effect when both causes are present is .75. These probabilities are captured in the $P(Y|X_1, X_2)$ column in Table 1.

Explaining away is the phenomenon that $P(x_i = 1|y = 1,\ x_j = 1) = 75/(75 + 50) = .60$ is lower than $P(x_i = 1|y = 1, x_j = 0) = 50/50 = 1.00$. Restated, X_1 and X_2 are negatively related within the subset of cases

[3] We use the term "explaining away" because an alternate label, "discounting", has many informal meanings and has been used in psychology to refer to other phenomena (Cheng & Novick, 2005, pp. 700–701; Khemlani & Oppenheimer, 2011, p. 2). In particular, "discounting" has often been used to refer to lowering one's estimate of the strength of one cause when one learns of a second cause that is strong (e.g., Goedert, Harsh, & Spellman, 2005), which is related to both rational and irrational forms of "cue competition," "blocking," and "conditioning." Note, the judgments assessed here are probability estimates, not causal strength judgments.

Table 1
Hypothetical "explaining away" data and inferences on $[X_1 \rightarrow Y \leftarrow X_2]$.

X_1	X_2	Y	Number of cases	$P(Y\|X_i, X_j)$	$P(X_i\|Y, X_j)$
1	1	1	75	$P(y = 1\|x_i = 1, x_j = 1) = 0.75$	$P(x_i = 1\|y = 1, x_j = 1) = 0.60$
1	1	0	25	$P(y = 1\|x_i = 1, x_j = 0) = 0.50$	$P(x_i = 1\|y = 1, x_j = 0) = 1.00$
1	0	1	50	$P(y = 1\|x_i = 0, x_j = 1) = 0.50$	$P(x_i = 1\|y = 0, x_j = 1) = 0.33$
1	0	0	50	$P(y = 1\|x_i = 0, x_j = 0) = 0.00$	$P(x_i = 1\|y = 0, x_j = 0) = 0.33$
0	1	1	50		
0	1	0	50		
0	0	1	0		
0	0	0	100		

Note: X_1 and X_2 can be thought of as two diseases that can each cause symptom Y. The number of cases column can be thought of as a tally of the number of patients with X_1, Y, and X_2.

when $y = 1$. The inference $P(x_i = 1|y = 1)$ is sometimes included in the phenomenon of explaining away; $P(x_i = 1|y = 1, x_j = 0) > P(x_i = 1|y = 1) > P(x_i = 1|y = 1, x_j = 1)$. The intuitive way to think about explaining away is that when there are two causes, X_1 and X_2, both of which are sufficient to produce $y = 1$, then when $y = 1$, if one of the causes is present, the other does not need to be present to explain why $y = 1$, but if one of the causes is absent, then the other must be present to explain why $y = 1$.

However, explaining away is subtle. Explaining away does not normatively occur when the effect is absent or in its "typical" state. For example, even though flu (X_1) and asthma (X_2) are negatively correlated among patients with a cough, among patients without a cough ($y = 0$) they are independent of one another. In the example in Table 1, both $P(x_i = 1|y = 0, x_j = 1) = 25/(25 + 50) = 1/3$ and $P(x_i = 1|y = 0, x_j = 0) = 50/(50 + 100) = 1/3$.

In the right column of Table 1, we list all four inferences of X_i given Y and X_j. When laid out this way it is easy to see that inferring X_i from Y and X_j is complicated. Whereas there is no effect of X_i on X_j within the subset when $y = 0$, there is a negative effect of X_i on X_j within the subset when $y = 1$. This means that when predicting X_i from Y and X_j, there is an interaction between Y and X_j. In contrast, predicting Y from X_i and X_j is simpler; increasing X_i and X_j always increases Y (see Table 1, $P(Y|X_1, X_2)$).

Additionally, explaining away is dependent on the exact manner in which the two causes combine to produce the effect. In the most common case described above when the integration follows a "Noisy-Or" function and the causes contribute independently and separately to the effect, then explaining away is the normative inference. However, explaining away is not the normative pattern when causes combine in other ways such as if they are both necessary for the effect (Rehder, 2015).

Despite the fact that explaining away has been regarded as a critical reasoning capability, and although social psychologists have long debated whether people explain away too much, too little, or appropriately (see McClure, 1998 for a review), in almost every instance in which explaining away has been studied in behavioral experiments, the parameters of the causal model were not precisely specified by the experimenters. Thus, the human judgments could not properly be compared to a normative standard.

Three general methods have been used to test if human reasoning is consistent with the normative standard for explaining away. The first approach involves telling participants a cover-story that implicitly sets up a causal structure with two causes and an effect, and having them make judgments that correspond to the parameters of the causal structure, including the base rates of the causes and the likelihood of the effect given each of the causes. Given a participant's parameters, it is possible to calculate what the explaining away inferences should be given his or her own beliefs. This is the approach that Morris and Larrick (1995) used and they found some explaining away though only about half the amount warranted by the normative model. One limitation on this approach is that there is little control over participants' beliefs about the parameters, and the normative amount of explaining away in this study turned out to be very small.

A second approach involves only testing for a directional effect via forced choice, whether people infer that $P(x_i = 1|y = 1, x_j = 0) > P(x_i = 1|y = 1, x_j = 1)$, which does not require presenting participants with the actual parameters. Rehder (2014) found trends towards explaining away in some experiments, but in others he found trends in the opposite direction. The weakness of this approach is that only qualitative, not quantitative explaining away judgments can be assessed.

A third method (Fernbach & Rehder, 2013, Experiment 3) involves giving participants the parameters of the causal structure via written instructions. Here again there is a trace of an explaining away effect, though much smaller than what would be expected normatively. The weakness of this approach is that the verbal statements of parameters may be an ineffective format to communicate information for probabilistic inferences.

In sum, the results from these three methods suggest that the explaining away principle is not generally respected. Explaining away occurs in some situations, but it is not clear what determines when it occurs; and when present, it under-estimates the normative effect.

In the present studies we tested for explaining away after participants had the opportunity to learn the statistical relationships from case-by-case experiences of the three variables in a common effect structure, as well as from written and graphical instructions. This approach circumvents many of the weaknesses of the prior approaches; we can set the parameters so that there normatively should be a large explaining away effect and we can examine both qualitative and quantitative judgments compared to the normative model.

In addition, presenting the statistical relations between the variables through case-by-case experience is most likely more effective than through verbal instruction. Previous research has shown that people are often better at performing probabilistic inference after experiencing the data than when the parameters are only stated verbally. In fact, people are notoriously bad at diagnostic inference tasks when the causal scenario is merely described verbally (Eddy, 1982; Kahneman & Tversky, 1973). Researchers have labeled the typical pattern of results as “base rate neglect.” A variety of factors can improve reasoning on diagnostic inference tasks so that people are sensitive to base rates, one of which is experiencing the statistical relationship between the two variables (Barbey & Sloman, 2007; Christensen-Szalanski & Beach, 1982; Koehler, 1996). People do not always respond perfectly when they have access to experience (Medin & Edelson, 1988; Reips & Waldmann, 2008), however most findings suggest that experience can help people make better diagnostic inferences, as well as other probabilistic judgments (Edgell et al., 2004; Hadar & Fox, 2009; Hertwig & Erev, 2009).

The question posed here is how accurately people perform explaining away after obtaining experience across representative learning trials with the appropriate statistical relationships between the three variables involved in the common effect structure. The previous literature, suggesting that experience can facilitate performance on diagnostic judgments, is relevant to this question because the diagnostic judgment, $P(X_i|Y)$, is one of the judgments involved in explaining away; explaining away also involves the inference $P(X_i|Y, X_j)$. Perhaps explaining away will be fairly accurate after participants have obtained appropriate experience. In Experiment 2, we tested explaining away with numerical variables. One specific reason for testing explaining away for numerical variables is that the explaining away phenomenon may be more intuitive when the two causes combine linearly (e.g., a sum or average) to produce the effect than when they combine through a noisy-OR probabilistic function (see Section 3.1.3 for more details). This is another reason why previous studies may have underestimated intuitive explaining away abilities.

1.4. Outline of current studies

We studied adherence to the Markov Assumption, the strengths of inferences, and explaining away inferences in two experiments, both of which provided participants with instructions concerning the causal network and direct exposure to the statistical parameters of the causal model through case-by-case learning trials. Experiment 1 investigated inference with binary variables with weaker (Experiment 1a) or stronger (Experiment 1b) parameters, and Experiment 2 investigated inference with numerical variables. Both experiments provided monetary incentives for correct responding.

2. Experiments 1a and 1b: Two studies of causal reasoning on events described as binary variables

2.1. Methods

2.1.1. Participants

Fifty-one (or 55 in Experiment 1b) undergraduates at the University of Chicago were paid approximately \$4 (\$6) to participate in a study that lasted 17 (32) min on average. To further motivate the participants, they were also paid 8 (10) cents for each correct inference.

2.1.2. Stimuli and design

Participants reasoned about three scenarios involving a chain [$X_1 \rightarrow Y \rightarrow X_2$], a common cause [$X_1 \leftarrow Y \rightarrow X_2$], and a common effect [$X_1 \rightarrow Y \leftarrow X_2$] structure. The variables (X_1, X_2, and Y) were framed as physiological variables in the human body that could be either high (represented as + or 1) or low (represented as − or 0). There were three cover stories, one about neurotransmitters (amounts of Serotonin, Epinephrine, and Dopamine), another about how the digestive tract absorbs chemicals from food (amounts of Water, Protein, and Fructose Absorption), and the last one about components of blood (Red Blood Cell, White Blood Cell, and Platelet Concentration). These variables were chosen so that they could plausibly be causally related to one another probabilistically in any possible combination, and participants would be very unlikely to have prior beliefs about how they were causally related. The order of the three causal structures, the cover-stories, the assignments of the three labels to variables (X_1, X_2, and Y), the position of the three variables on the computer screen, and the order of the learning trials were all randomized.

The sets of learning trials (Table 2) were generated in the following way: For the chain and common cause structure, the chosen parameters produced identical sets of learning trials. The base rates of all three variables were .5; the variables were equally likely to be present or absent. The difference between Experiments 1a and 1b was the strength of the causes. When a cause was present it produced its effect with probability .75 in Experiment 1a and with probability .875 in Experiment 1b. When a cause was absent its effect still occurred with probabilities of .25 and .125 in Experiments 1a and 1b respectively. These manipulations meant that inferences such as a transitive inference $P(x_i = 1|x_j = 1)$ should be more extreme (stronger) in Experiment 1b.

For the common effect structure, the difference between Experiment 1a and 1b was that the base rates of the two causes, $P(x_i = 1)$ and $P(x_j = 1)$ were .50 in Experiment 1a and .25 in Experiment 1b. This manipulation implies a stronger normative explaining away effect in Experiment 1b. In both experiments, the two causes combined through a Noisy-OR gate (Pearl, 1988) with strengths of .50, and thus $P(y = 1|x_1 = 0,\ x_2 = 0) = 0$, $P(y = 1|x_1 = 1,\ x_2 = 0) = P(y = 1|x_1 = 0,\ x_2 = 1) = .50$, and $P(y = 1|x_1 = 1,\ x_2 = 1) = .75$. This meant that the base rate for the effect, $P(y = 1)$ was .43 for Experiment 1, and .23 for Experiment 1b.

These parameters were chosen to be theoretically neutral – the causal strengths were moderately strong (not extremely strong or extremely weak), and the causes were chosen to have base rates in the middle of the scale, not to be very common or very rare. These parameters can also be instantiated in fairly few learning trials; using more extreme parameters would require many more learning trials.

The normative point estimate inferences in Table 3 can be computed directly from Table 2. For example, $P(y = 1|x_1 = 1)$ can be computed by dividing the sum of all the rows in which $y = 1$ and $x_1 = 1$ by the sum in all the rows in which $y = 1$ (e.g., $[9 + 3]/[9 + 3 + 3 + 1] = .75$ in Experiment 1a). Alternatively, if people learn the parameters of the causal model from experience the normative inferences can be derived from the parameters.

2.1.3. Procedures

The general procedure followed a standard trial-by-trial, case-by-case causal learning paradigm in which participants were first told a causal cover-story, then learned the probabilistic relations between the variables from experience, and finally made a series of inferences and judgments.

Participants were asked to pretend that they were physiologists studying biological processes in the human body. They were told that they were performing studies in which they would bring healthy

Table 2
Learning trials in Experiments 1a and 1b.

X_1	Y	X_2	Number of trials			
			Chain and common cause		Common effect	
			Exp. 1a	Exp. 1b	Exp. 1a	Exp. 1b
1	1	1	9	49	6	6
1	1	0	3	7	4	12
1	0	1	1	1	2	2
1	0	0	3	7	4	12
0	1	1	3	7	4	12
0	1	0	1	1	0	0
0	0	1	3	7	4	12
0	0	0	9	49	8	72

Table 3
Key inference questions by causal structure in Experiments 1a and 1b.

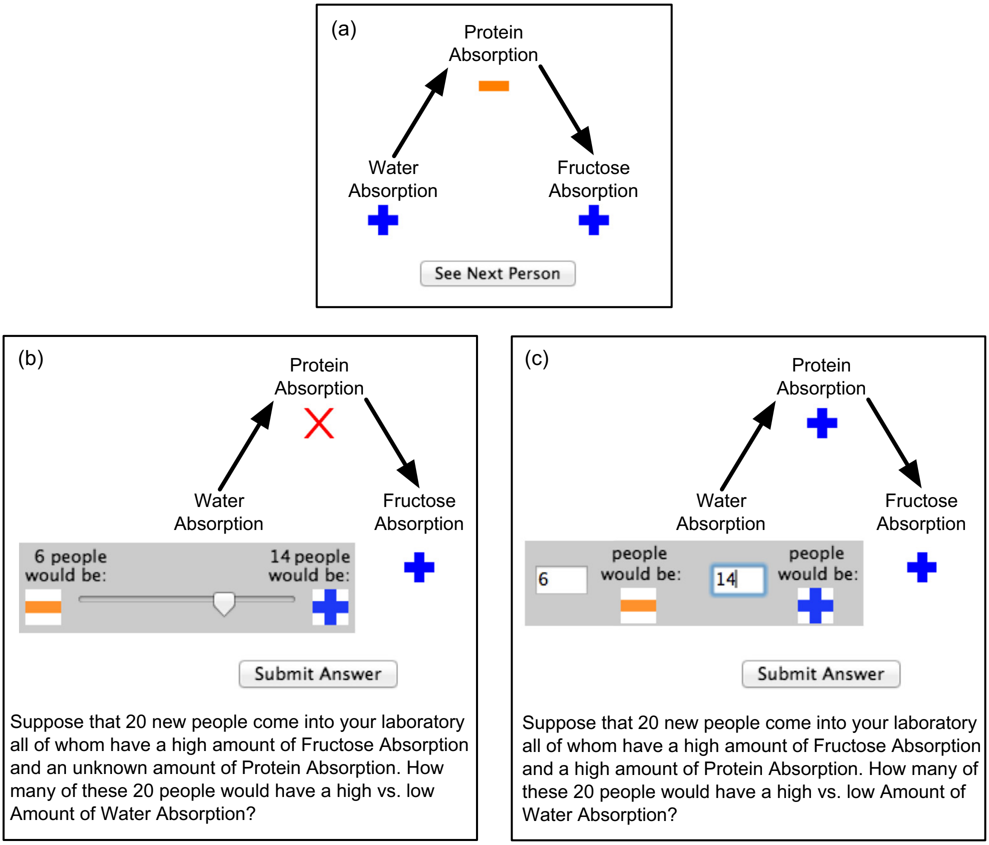
Inference type	Specific inferences	Normative judgments	
		Exp. 1a	Exp. 1b
Chain and common cause			
Markov assumption	$P(x_i = 1 \mid y = 1, x_j = 1)$, $P(x_i = 1 \mid y = 1, x_j = 0)$	.75	.875
	$P(x_i = 1 \mid y = 0, x_j = 0)$, $P(x_i = 1 \mid y = 0, x_j = 1)$	.25[*]	.125[*]
Transitive inferences	$P(x_i = 1 \mid x_j = 1)$	.625	.78
	$P(x_i = 1 \mid x_j = 0)$	.375[*]	.22[*]
Inferences of the middle variable	$P(y = 1 \mid x_i = 1, x_j = 1)$	.90	.98
	$P(y = 1 \mid x_i = 0, x_j = 0)$	.10[*]	.02[*]
Common effect			
Markov Assumption	$P(x_i = 1 \mid x_j = 1)$, $P(x_i = 1 \mid x_j = 0)$	.50	.25
Explaining away	$P(x_i = 1 \mid y = 1, x_j = 1)$	.60	.33
	$P(x_i = 1 \mid y = 1)$	.71	.60
	$P(x_i = 1 \mid y = 1, x_j = 0)$	1	1

[*] For simplicity of exposition, these inferences were flipped to the upper portion of the probability scale for analysis. For example, .25 was converted to .75.

people into a laboratory and would measure three physiological variables. They were told how to interpret pictures like the ones in Fig. 1a, where arrows represented causal relations between the three physiological variables and pointed from causes to effects. A "+" sign signified a high amount of the variable and "−" a low amount of the variable. Participants were also told not to use any prior knowledge about physiology and to assume that these three variables are the only ones that mattered within this biological system.

Next, participants completed a learning phase for each of the three causal scenarios in a randomized order, involving a chain, common cause, and common effect. Participants were shown a graphical representation of the causal relationships (Fig. 1a) and they observed whether each of the variables was high or low in a sample of 32 (128) cases ("healthy people"). The cases were presented in a sequential trial-by-trial format in a randomized order and the positions of the three variables, X_1, X_2, and Y on the screen were randomized.

After the learning phase participants made a series of inferences; the order of the questions was randomized. Participants made inferences about each variable, given that the states of the other two variables were high, low, or unknown. Table 3 shows the key inferences that pertain to specific questions about the Markov Assumption, the strength of transitive inferences and inferences on the middle variable, and explaining away. The questions were presented to participants using both a visual diagram and corresponding text (see Fig. 1b and c). When the state of a variable was unknown it was denoted visually with an X mark (see Fig. 1b) and participants were told that the machine used to test for that variable was broken.



Suppose that 20 new people come into your laboratory all of whom have a high amount of Fructose Absorption and an unknown amount of Protein Absorption. How many of these 20 people would have a high vs. low Amount of Water Absorption?

Suppose that 20 new people come into your laboratory all of whom have a high amount of Fructose Absorption and a high amount of Protein Absorption. How many of these 20 people would have a high vs. low Amount of Water Absorption?

Fig. 1. Example stimuli in Experiment 1. Note: Panel (A) shows an example of one trial in the learning phase. Panels (B and C) show examples of how participants made two inferences: $P(\text{water} = 1|\text{fructose} = 1)$ and $P(\text{water} = 1|\text{fructose} = 1, \text{protein} = 1)$. Panels (B and C) show the two alternate input methods (slider and text boxes) used in Experiments 1a and 1b respectively.

Underneath the variable to be inferred was a gray box that participants used to input their estimates. Following the practice of Waldmann and Hagmayer (2005), we used a frequency format (number of people out of 20) for the question rather than a probability format. Participants submitted their responses using either a slider (Fig. 1b, Experiment 1a) or two text boxes – after typing a response (e.g., 14) in one text box, the computer immediately displayed the result (e.g., 20 – 14 = 6) in the other (Fig. 1c, Experiment 1b). We used these two methods to verify that the results were not response-scale dependent and also to avoid the possibility of anchoring on the middle response with the slider.

At the end of the study, participants were paid for their time and a bonus for the number of questions that they answered correctly; an answer was considered correct if it was either exactly correct according to the normative calculation or if the chosen value was as close as could be obtained on the 21-point scale. Participants were not given feedback during the test phase of the experiment.

2.2. Analyses

The following analytical strategies hold for all the experiments. All responses were converted to a probability scale of 0–1. Because a common cause $[X_1 \leftarrow Y \rightarrow X_2]$ is symmetric, inferences like $P(x_1 = 1|x_2 = 1)$ and $P(x_2 = 1|x_1 = 1)$ are essentially duplicates, so they are labeled $P(x_i = 1|x_j = 1)$ and treated as repeated measures. For the chain $[X_1 \rightarrow Y \rightarrow X_2]$, we looked and did not find systematic differences for the inferences going down versus up the chain (see Appendix A), thus we treat inferences like $P(x_1 = 1|x_2 = 1)$ and $P(x_2 = 1|x_1 = 1)$ as repeated measures. When the inferences were symmetric we converted inferences in the bottom half of the scale to the top half so they could be analyzed together like $P(x_i = 1|x_j = 0)$ and $P(x_i = 1|x_j = 1)$, or $P(x_i = 1|y = 0, x_j = 0)$ and $P(x_i = 1|y = 1, x_j = 1)$ for the chain and common cause; see Table 3. These inferences were symmetric and analyzing them together simplifies reporting the results dramatically.

2.3. Results and discussion

2.3.1. Success of the standard model

Most of the previous work has concluded that people get many of the inferences approximately right. The current study agrees with this assessment. Fig. 2 plots all the inferences and their normative answers in Experiment 1; ideally responses should fall on the diagonal. Correlations between normative and human responses are $r^2 = 0.47$ for Experiment 1, meaning that the standard normative model does explain sizeable amounts of variance. The following sections focus on the ways that the judgments deviate from the normative model.

2.3.2. Violations of the Markov Assumption

The Markov Assumption implies that pairs of inferences such as $P(x_i = 1|y = 1, x_j = 1)$ and $P(x_i = 1|y = 1, x_j = 0)$ on the chain and common cause should be equivalent. Fig. 3 is a summary display of distributions of individual judgments in Experiments 1a and 1b for the chain and common-cause networks. The "high" and "low" judgments map onto X_j, the screened off variable being 1 versus 0. The first impression on seeing these graphs is that there is enormous variation in participants' inferences for questions that have exact normative "answers." Second, judgments varied systematically depending on the state of X_j; the judgments appear to be higher for the "high" $P(x_i = 1|y = 1, x_j = 1)$ than the "low" $P(x_i = 1|y = 1, x_j = 0)$ inferences. This is driven, to a large extent, by subjects often judging $P(x_i = 1|y = 1, x_j = 0)$ to be exactly 0.50. This response tendency makes sense if participants believe that both X_j and Y are relevant for inferring X_i, and weight the two equally.

However, there are also a number of ways that subjects are clearly sensitive to the prescriptions of the normative model. The inferences were (properly) insensitive to whether the causal structure was a chain or a common-cause network. Additionally, participants were sensitive to variations in the causal strengths communicated through the case-by-case learning experiences. The distributions of responses (properly) are higher for Experiment 1b than 1a.

To test whether the average judgments were reliably different for the $P(x_i = 1|y = 1, x_j = 1)$ versus $P(x_i = 1|y = 1, x_j = 0)$ judgments, we used mixed linear regressions. When appropriate, we used a negative square root transformation on the dependent variable to transform the data to rough normality. All confidence intervals reported were back-transformed so that they can be interpreted on the probability scale.[4] For one participant in Experiment 1a the participant's responses were very similar within a scenario, likely reflecting disengagement from the task. Thus, we threw out those observations in order not to bias the results towards weak inferences.

We ran four mixed effects regressions for the chain and common cause, and for Experiment 1a and 1b, to test whether the inferences were higher when the screened-off variable (X_j in Table 4) was 1 instead of 0. By-subject random effects were included for the intercept and for the slope (the difference between the two inferences). See Table 4 for 95% confidence intervals on the size of the Markov violation; three of the violations were significant, and the violation for the common cause in Experiment 1b was nearly significant.

[4] Because it was not always possible to transform the data to normal distributions, all analyses were also run using mixed effect logistic regressions using median splits. For example to compare the inferences $P(x_i = 1|y = 1, x_j = 1)$ and $P(x_i = 1|y = 1, x_j = 0)$, we took the median judgment across both types of inference, recoded inferences above the median as 1, below as 0, and then used a logistic regression with the same random effects structure as above to test whether $P(x_i = 1|y = 1, x_j = 1)$ was more likely than $P(x = 1|y = 1, x_j = 0)$ to have 1 s. These two methods produced very similar results, if anything the median split analysis tended to produce cleaner results. However, we report the normal regressions because they can be back-transformed onto the probability scale which aids interpretability.

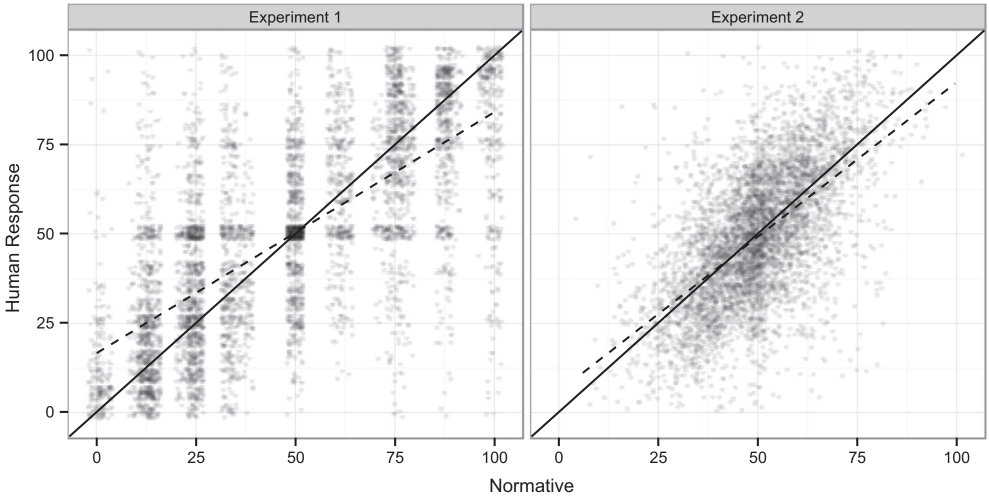


Fig. 2. Human responses versus normative answers.

The Markov Assumption also has a role in the common effect [$X_1 \rightarrow Y \leftarrow X_2$] structure; X_1 and X_2 are independent of each other when the state of Y is not known, which means that $P(x_i = 1|x_j = 1) = P(x_i = 1|x_j = 0)$. We compared the difference between these two inferences. In Experiment 1a there was essentially no difference; no violation of the Markov Assumption. However, in Experiment 1b there was another significant violation of the Markov Assumption (see Table 5).

We also assessed whether a small minority of participants were responsible for the Markov violations, or whether violating the Markov Assumption was a common habit. Each participant made 12 inferences relevant to the Markov Assumption; see Table 3. For each participant we conducted a *t*-test comparing the 6 inferences when the irrelevant variable was 1 against the six inferences when the irrelevant variable was 0. Out of a total of 106 participants, 87 gave higher inferences when the irrelevant variable was 1 than 0, and for 32 participants this effect was significant (despite the fact that each *t*-test was computed with only 12 judgments). If there really is no overall tendency to violate the Markov Assumption, given a bidirectional $\alpha = .05$, only about 3 participants should have a significant positive Markov violation merely due to chance. In sum, the habit to violate the Markov Assumption appears to be common.

2.3.3. The strength of inferences

We were interested in whether the transitive and middle inferences were normatively strong or whether they were too weak (too close to the base rate of .50). Fig. 4 provides graphs of the distributions of responses to questions assessing transitive inference strength; the vertical bars give the normative answers. Table 6 also presents the means and the normative answers. As in our examination of adherence to the Markov Assumption (Fig. 3), the first impression on seeing these graphs is that there is enormous variability across participants. Many distributions have groups of responses both considerably above and below the normative, and some have small tails of completely unreasonable responses below .50. Second, as in the Markov Violation inferences, there is sensitivity to the underlying probability values, with a (proper) shift upwards when the normative answer was higher in Experiment 1b. For simple transitive inferences in Experiment 1a, where the normative answer is .625, the average ratings were close to correct, but for each structure, there is a large spike of responses at .50 (the base rate), likely reflecting a rounding habit given that the normative judgments are only slightly above .50. In Experiment 1b, where the normative response should be .78, the average response is too low and, again, there are spikes of responses at .50 (most dramatically for the common cause structure).

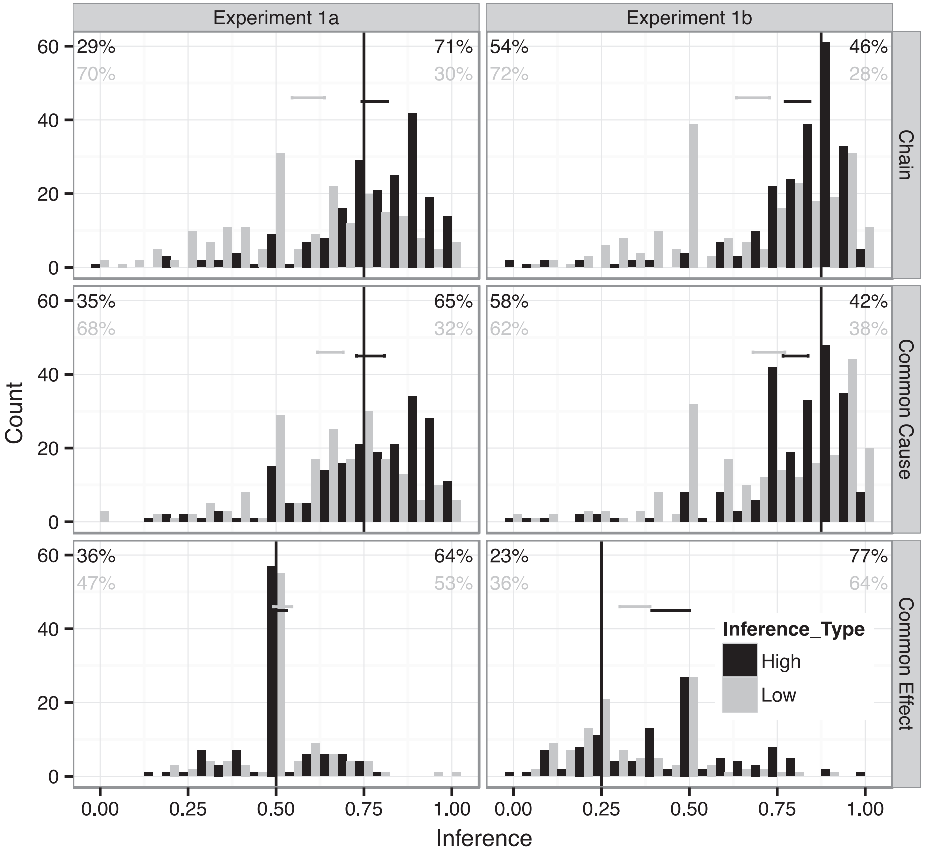


Fig. 3. Distributions of individual responses showing Markov Violations from Experiment 1. Note: Gray bars indicate responses when the screened-off variable has a low value, darker bars when the screened-off variable has a high value (the lighter and darker bars should be identical, if the Markov Assumption holds). The thin vertical line in each panel represents the 'correct' normative point-estimate response. The horizontal bars high up in each panel represent 95% confidence intervals on the means for each condition within the panel computed from regressions with by-subject random effects on the intercept. The numbers are the percent of inferences on either side of the normative calculation, after removing judgments that are exactly correct.

Table 4
Tests of the Markov Assumption in Experiment 1a and 1b for chain and common cause.

Inference	Norm.	$X_1 \rightarrow Y \rightarrow X_2$		$X_1 \leftarrow Y \rightarrow X_2$	
		Mean	%>Norm	Mean	%>Norm
Exp. 1a					
$P(x_i = 1 \mid y = 1, x_j = 1)$	.75	.78	71%**	.77	65%*
$P(x_i = 1 \mid y = 1, x_j = 0)$	.75	.59	30%**	.65	32%**
95% CI of difference	0	[.13, .22]	–	[.07, .16]	–
Exp. 1b					
$P(x_i = 1 \mid y = 1, x_j = 1)$	.875	.81	46%	.80	42%
$P(x_i = 1 \mid y = 1, x_j = 0)$	.875	.68	28%**	.73	38%*
95% CI of difference	0	[.06, .15]	–	[−.01, .10]	–

Note: For the percent in the %>Norm, %<Norm columns, we dropped all inferences that were exactly equal to the normative answer.

[*] $p \leqslant .05$.

[**] $p \leqslant .01$.

Table 5
Tests of the Markov Assumption in Experiment 1a and 1b for common effect.

Inference	Norm.	Mean	%>Norm
Experiment 1a			
$P(x_i = 1 \mid x_j = 1)$	.50	.52	64%†
$P(x_i = 1 \mid x_j = 0)$	.50	.51	53%
95% CI of difference	0	[−.02, .05]	–
Experiment 1b			
$P(x_i = 1 \mid x_j = 1)$	.25	.45	77%**
$P(x_i = 1 \mid x_j = 0)$	.25	.35	64%
95% CI of difference	0	[.04, .16]	–

Note: For the percent in the >Norm column, we dropped all inferences that were exactly equal to the normative.
* $p \leqslant .05$.
** $p \leqslant .01$.
† $p = .06$.

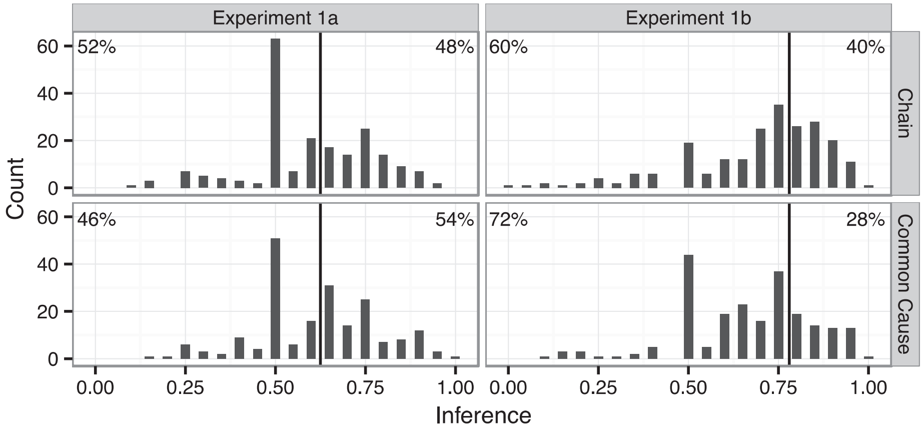


Fig. 4. Distributions of responses to questions testing transitive inferences in Experiment 1. Note: The thin vertical line in each panel represents the 'correct' point-estimate normative response. The numbers are the percent of inferences on either side of the normative calculation.

Table 6
Strength of transitive and middle inferences (means and % responses > normative) in Experiment 1.

Inference	Norm.	$X_1 \rightarrow Y \rightarrow X_2$		$X_1 \leftarrow Y \rightarrow X_2$		$X_1 \rightarrow Y \leftarrow X_2$		
		M	%>Norm.	*M*	%>Norm.	Norm.	*M*	%>Norm.
$P(x_i = 1 \mid x_j = 1)$; trans. Exp. 1a	.625	.60	48%	.62	54%	–	–	–
$P(x_i = 1 \mid x_j = 1)$; trans. Exp. 1b	.78	.69	40%†	.67	28%**	–	–	–
$P(y = 1 \mid x_i = 1, x_j = 1)$; middle Exp. 1a	.90	.80	25%**	.82	35%**	.75	.79	74%**
$P(y = 1 \mid x_i = 1, x_j = 1)$; middle Exp. 1b	.98	.86	12%**	.88	15%**	.75	.75	76%**

Note: Norm. = Normative. Trans. = Transitive. For the percent in the %Norm columns, we dropped all inferences that were exactly equal to the normative answer. Transitive inferences for the common effect are discussed in the Markov Violations section.
* $p \leqslant .05$.
** $p \leqslant .01$.
† $p = .07$.

Because some of the distributions of the judgments were skewed, we performed non-parametric statistics to test whether the inferences, in general, were too strong or too weak.[5] For inferential statistics, we recoded the inferences as larger (1), equal to (0), or smaller (−1) than the normative inference. Then we took the average of these scores within a participant, and compared all the averages against 0 using a Wilcoxon test. This is essentially a test of whether the median response is different from the normative calculation. We also eliminated judgments that were extremely low, lower than 1 minus the normative value from the analysis.[6] For descriptive statistics, we report the mean on the probability scale (0–1), and we also report the percent of inferences that are larger than the normative inference (dropping all inferences that are exactly equal to the normative inference). If participants' inferences were appropriately strong then 50% of the judgments would be higher and 50% lower than the normative answer.

Table 6 presents the means of the inferences, the percentages of inferences above the normative calculation (asterisks indicate whether the responses were significantly different from the normative calculation according to the Wilcoxon test). The transitive inferences were too conservative in 1b; they were not significantly different from normative in Experiment 1a. Note that Experiment 1a does not provide a strong test for the strength of transitive inferences because the normative answer was .625, close to the middle of the scale. This means it would be hard to detect a conservative pattern in Experiment 1a. But, the normative answer in Experiment 1b was .78, providing a more powerful test of the weak inference hypothesis.

The inferences about the values of the "middle variable" (Table 6, Fig. 5) were too conservative in Experiment 1a, for Chain and Common-Cause structures, although most inferences are above 0.75 and so do not show the conservative habit of anchoring on the base rate (.50) seen for transitive inferences. In Experiment 1b the normative inference is .98, and technically these inferences are also too weak. However, looking at the distributions the most common answer was 0.95, so participants are for the most close to correct.

Common-Effect middle-variable inferences revert to non-normative variability with many apparently uninformed responses. For high-valued inferences (where both causes of the common effect are present) many judgments over-shoot the normative answer. Responses are more sensible for the low-valued inferences: most participants infer the effect will not occur; but there is another long tail of uninformed responses.

Tables 4 and 5 also present the percent of Markov Assumption inferences greater than and less than normative, after removing any inferences that are exactly normative. This metric gives an easy way to understand whether the judgments are too strong or too weak. As would be expected from the other analysis, these judgments tend to be too high in the high conditions (though not in Experiment 1b, when the normative inference is very high), and too low in most of the low conditions (though not for the Common Effect conditions).

2.3.4. Explaining away

Distributions of individual responses, relevant to the test of explaining away are presented in Fig. 6. The normative pattern is that $P(x_i = 1|y = 1,\ x_j = 0) > P(x_i = 1|y = 1) > P(x_i = 1|y = 1,\ x_j = 1)$. Our first impression on seeing these graphs is that out of all the inferences in Experiment 1, these inferences show the highest variability in responses, despite the fact that they still have sharply defined normative answers. Second, whereas the other inferences generally moved in the right direction when comparing different types of inferences and different parameters, explaining away judgments do not track the directional patterns of the normative inferences.

[5] The previous tests of the violations of the Markov Assumption used transformations and parametric statistics. However, this approach does not make sense when comparing a skewed distribution against a single point-estimate normative value because any transformation shifts the mean. To be conservative, we used non-parametric statistics.

[6] We were worried that participants might sometimes accidentally enter a response using the wrong end of the scale. For example if participants just answered $P(x_i = 1|x_j = 1) = .70$, when faced with answering $P(x_i = 1|x_j = 0)$ they might think that the inference should be .20 away from the middle of the scale but forget to flip the response to .30 and instead type in .70. Obvious mistakes that are clearly on the opposite side of the scale were fairly rare (see the figures, especially Fig. 4, which included all the data), but to be conservative we eliminated extreme data on the opposite side of the scale from the normative inference.

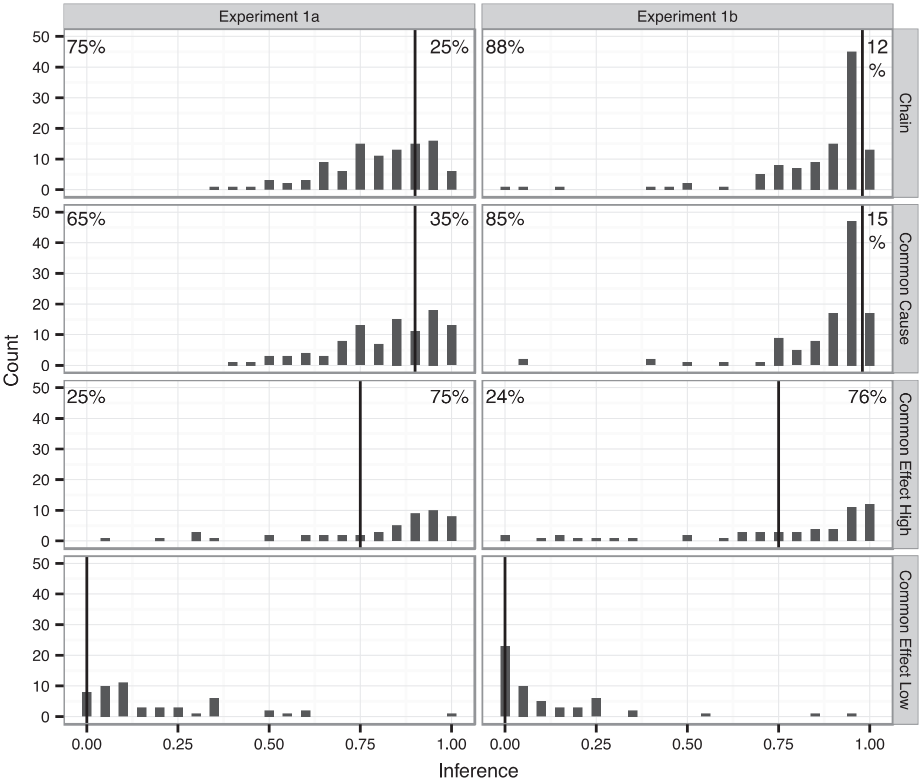


Fig. 5. Inferences to "middle" variables in Experiment 1. Note: The normative common effect low $P(y = 1|x_i = 0, x_j = 0)$ inferences are not equal to the normative common effect high $P(y = 1|x_i = 1, x_j = 1)$ inferences, so they are not flipped to the upper end of the scale and are presented separately. The thin vertical line in each panel represents the 'correct' normative response. The numbers are the percent of inferences on either side of the normative calculation, after removing judgments that are exactly correct.

Our reading of the results is that in only in the "alternate cause did not occur" condition, did a substantial number of participants show a grasp of the relevant normative principle: There is a spike of participants with the correct answer (1.00) in Experiment 1b, for the $P(x_i = 1|y=, x_j = 0)$. That is, when there were only two possible causes for an effect that did occur, and one cause did not occur, 32 responses (out of 110) concluded (properly) that the other cause must have occurred. But, note that in Experiment 1a, fewer than 10 responses were correct for the same inference.

The left side of Table 7 shows the normative calculations and empirical means for the six inferences. The right side shows confidence intervals of the difference of means such as $P(x_i = 1|y = 1) - P(x_i = 1|y = 1, x_j = 1)$ that provide the crucial tests of explaining away. The confidence intervals were calculated using mixed linear regressions with by-subject random effects on the intercept and the slope (the difference between the two judgments) to account for repeated measures. The lower bound of the confidence interval identifies whether the amount of explaining away is significantly higher than zero and the upper bound identifies whether the amount of explaining away is significantly lower than the normative amount.

In Experiment 1a, which used base rates of .50, the inferences for $P(x_i = 1|y = 1)$ were on average lower than the inferences for $P(x_i = 1|y = 1, x_j = 1)$, not higher as implied by the normative model. And, the salient .50 base rate serves as an anchor for substantial numbers of participants (Fig. 6). The inferences for $P(x_i = 1|y = 1, x_j = 0)$ and $P(x_i = 1|y = 1, x_j = 1)$ were not significantly different.

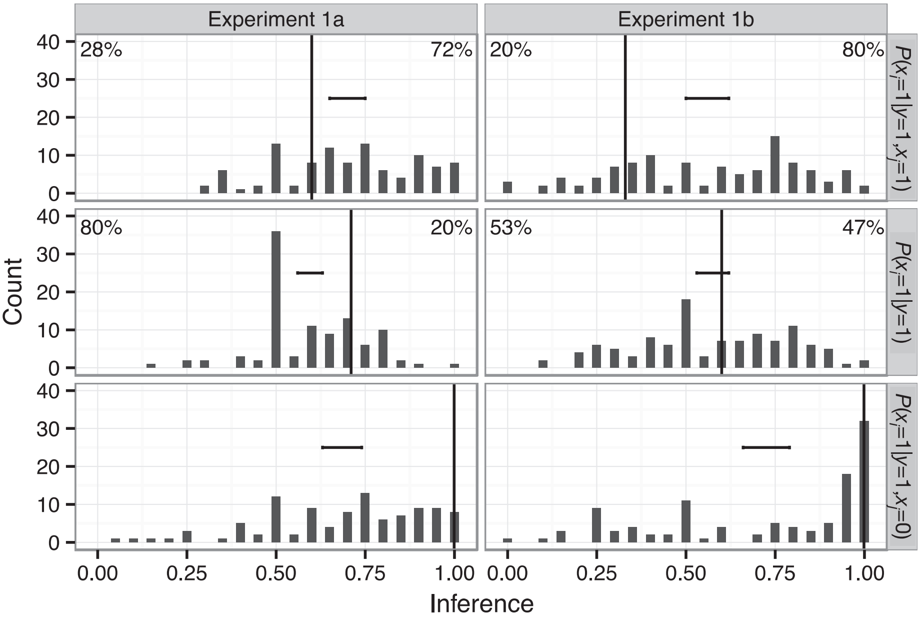


Fig. 6. Distributions of individual responses that tested “explaining away” in Experiment 1 on common effect [$X_1 \rightarrow Y \leftarrow X_2$]. Note: Vertical lines are the normative answers. Horizontal bars are 95% CI of the mean inference. The numbers are the percent of inferences on either side of the normative calculation, after removing judgments that are exactly correct.

In Experiment 1b, which used base rates of .25. The inferences of $P(x_i = 1|y = 1)$ and $P(x_i = 1|y = 1, x_j = 1)$ were not significantly different. The inferences for $P(x_i = 1|y = 1, x_j = 0)$ were significantly higher than $P(x_i = 1|y = 1, x_j = 1)$, though the difference was smaller than the normative calculation.[7]

In sum, we found no explaining away with base rates of .50 and we found some, but insufficient explaining away with base rates of .25.

2.3.5. Summary

In sum, Experiment 1 consistently found violations of the Markov Assumption and insufficient explaining away. There was an overall tendency for the inferences to be too weak; but still some were too strong (e.g., the inferences on the effect variable in the common effect structure when both causes were present and some of the inferences relevant to the Markov Assumption when both the relevant and irrelevant variables were present). These results are generally consistent with the conclusions from Rottman and Hastie (2014) based on a review of the literature.

[7] We also tested if there might be order effects in the explaining away judgments. In the chain and common cause structures, effects occurred even when their causes did not, but for the common effect structure, the effect only occurred if at least one cause occurred. If participants somehow transferred their parameter beliefs from the chain or common cause to the common effect condition, the amount of explaining away could appear smaller than it actually is. (We thank an anonymous reviewer for suggesting this possibility.) To test this possibility, we subtracted judgments of $P(x_i = 1|y = 1, x_j = 1)$ from $P(x_i = 1|y = 1, x_j = 0)$, to compute a single measure of the strength of explaining away. We then compared this strength of explaining away for participants who received the common effect structure first versus second or third. For Experiment 1a, the mean size of the discounting strength was not different when the common effect structure was first ($M = .04$, $SD = .22$), or second or third ($M = -.03$, $SD = .22$), $t(48) < 1$, $p = .36$. For Experiment 1b, the mean size of the discounting strength was larger when the common effect structure was first ($M = .29$, $SD = .28$) compared to second and third ($M = .12$, $SD = .26$), $t(53) = 2.10$, $p = .04$. However, even when the common effect structure was first, the magnitude of the explaining away effect was still less than half of the normative amount of .67, $t(14) = 5.12$, $p < .001$.

Table 7
Explaining away results from Experiment 1 on common effect [$X_1 \rightarrow Y \leftarrow X_2$].

Raw inferences			Explaining away comparisons	
Inferences	Norm.	Emp.	Norm.	95% CI of difference
Experiment 1a: base rates = .50				
$P(x_i = 1 \mid y = 1, x_j = 1)$	.60	.70	–	–
$P(x_i = 1 \mid y = 1)$	.71	.59	.11	[−.16, −.06]
$P(x_i = 1 \mid y = 1, x_j = 0)$	1	.69	.40	[−.08, .06]
Experiment 1b: base rates = .25				
$P(x_i = 1 \mid y = 1, x_j = 1)$	.33	.56	–	–
$P(x_i = 1 \mid y = 1)$	.60	.58	.27	[−.04, .07]
$P(x_i = 1 \mid y = 1, x_j = 0)$	1	.73	.67	[.09, .23]

Note: Norm. = normative. Emp. = empirical mean. The right half of the table reports the difference of the 2nd and 3rd rows to the top row.

3. Experiment 2: Causal reasoning on events described as numerical variables

3.1. Motivation

People frequently reason about magnitudes (e.g., temperature, intensity of back pain, score on an exam, speed of a car, degrees of happiness) instead of merely binary values. But, the vast majority of scientific studies of causal reasoning have relied on binary events. The primary motivation for Experiment 2 is to test the normativity of causal inferences in a scenario with numerical magnitudes using a format based on the materials and procedure in Experiment 1 (with binary variables). The normative model we use is linear regression, which has also been applied extensively as a statistical tool for scientific data analysis and as a normative model of human inference (Dawes & Corrigan, 1974; Hogarth & Karelaia, 2007; Kahneman & Tversky, 1973). A Bayesian causal network can be viewed as a collection of regression models; each effect is modeled as the outcome variable with all of its direct causes as predictors (Heckerman, 1998).[8]

A review of the sparse prior research on causal reasoning about numerical variables does not support confident hypotheses about whether we will see more or less rationality in Experiment 2, as compared to Experiment 1.

3.1.1. Markov Assumption

One reason people might violate the Markov Assumption is because they believe that the variables are not perfectly observed when they are presented in a "coarse" binary manner (Rehder & Burnett, 2005, called this the "uncertainty model"). If uncertainty concerning the exact states of variables that are expressed "coarsely" in binary values could produce apparent Markov Violations, then numerical expressions of variables should be perceived as more definite and certain, and violations of the Markov Assumption would be diminished. Consider the chain [$X_1 \rightarrow Y \rightarrow X_2$]. Suppose you are told that y is present but X_2 is absent and you are asked to infer X_1. Suppose further that you believe that X_1, Y, and X_2 can actually assume any state from 0 to 100, and "present" refers to a value greater than or equal to 50 and "absent" refers to a value less than 50. Given that the binary states of Y and X_2 conflict (y = present, but x_2 = absent), one might presume that both y and x_2 are close to 50. In that case one might infer that X_1 is also fairly close to 50. However, if you are told that y = present and x_2 = present, you might assume that they are both strongly present (e.g., maybe somewhere near 75), and then infer that X is strongly present. In summary, if people view binary variables as coarse simplifications of

[8] It is not possible to make exact comparisons between reasoning with binary and numerical variables for a variety of reasons. First, the function defining how multiple causes combine to produce an effect is necessarily different; the most typical functions are the Noisy-OR for binary variables and a linear, additive function for numerical variables. Second, in the numerical case it does not make sense to ask participants to make inferences such as $P(x_1 = 1 \mid x_2 = 1)$ and $P(x_1 = 1 \mid x_2 = 0)$, but rather it is appropriate to ask about the expected value of X_1 given a range of values of X_2. Third, even if we asked participants to make the same inferences as in the binary case such as $P(x_1 = 1 \mid x_2 = 1)$ and $P(x_1 = 1 \mid x_2 = 0)$, the normative answers would be different due to the different mathematical integration functions.

variables that are actually magnitudes, the most plausible hypothesis is that people will be more likely to respect the Markov Assumption when reasoning about magnitude variables.

3.1.2. Strength of inferences

As explained in the introduction, some previous studies (Kahneman & Tversky, 1973) have found that people are non-regressive and provide a response that matches the magnitude (or extremity) of a given variable. In the current study which used variables on a 0–100 scale with means of 50, a non-regressive response for $E(X_i|x_j = 75)$ would be answering 75 or higher, when the normative answer is 67 (E stands for expectation). In this case, people's inferences would be too strong. Kahneman and Tversky cited their representativeness heuristic as an explanation for this pattern of non-regressive judgments. Another way to explain such a finding would be through anchoring on the predictor, 75, and insufficiently adjusting towards 50. We hypothesized that in Experiment 2, the relatively barren materials and the numerical stimuli in these experiments, could likely evoke an anchor-and-adjust heuristic strategy. Our best speculation is that a similar anchor-and-adjust process will describe many participants' inference processes. The open question is what values will be selected as anchors in our Experiment 2. Participants could anchor on the salient cue value(s), resulting in inferences that are too strong. Alternately, they might anchor on the base rate value (50), which is likely to be reinforced by our learning from exemplars procedure, resulting in weak inferences.

3.1.3. Explaining away

Experiment 2 tests whether people explain away appropriately when they have learning experiences with cases in which two numerical causes combine to produce the effect across several trials. Nisbett and Ross suggested that a simple "hydraulic heuristic" was relied on in some circumstances, "as if causal candidates competed with one another in a zero-sum game" (1980, p. 128). For example, suppose that for the common effect structure $[X_1 \rightarrow Y \leftarrow X_2]$, $Y = X_1 + X_2$. If we know that $y = 10$ and $x_2 = 7$, we would infer that $x_1 = 3$. When the value of Y is known, the higher that X_2 is, the lower X_1 must be. This negative dependency also occurs for other functions such as an average (cf. Anderson, 1981) and for other linear functions. Of course, if there is noise in the system (Y is the sum of X_1, X_2, plus noise), then the relationship between X_1 and X_2 holding Y constant is not perfectly one-to-one, though they would still be negatively related.

Although these speculations are plausible (we too have an intuition that under some circumstances a hydraulic heuristic may be followed), Nisbett & Ross did not actually report on experiments that demonstrated an explaining away result. The only study that we know that has investigated explaining away with continuous variables, found either no or weak explaining away, though participants were not told how X_1 and X_2 combined to produce Y (e.g., $Y = X_1 + X_2$), so there is not a normative value against which to compare the human judgments (Sussman & Oppenheimer, 2011). Thus, we do not have a clear prediction as to whether explaining away is more likely to be respected when thinking about numerical magnitudes, rather than binary variables.

In sum, the main purpose of Experiment 2 is to describe human causal inference on three-variable causal structures with numerical magnitudes. For each of the three main deviations from normality seen with binary variables, violations of the Markov Assumption, weak inferences, and failures of explaining away, we speculated about why normative inference may be easier with magnitudes as opposed to binary variables. But, there is no evidence to support strong predictions one way or the other. Furthermore, because it is impossible to directly compare reasoning with binary versus numerical variables, the goal of Experiment 2 is mainly to obtain a precise description of human causal reasoning with numerical magnitude variables.

3.2. Methods

3.2.1. Participants

Fifty undergraduates at the University of Chicago were paid $12 per hour to participate in a study that lasted 32 min on average. They were also paid 10 cents for each judgment accurate within 3 points on either side of the correct response.

3.2.2. Stimuli and design

We used the same cover-story as the previous experiments, and a procedure with 32 learning trials. All three variables, X_1, X_2, and Y, were normally distributed with a mean of 50 and a standard deviation of 20 (constrained such that the minimum and maximum were 0 and 100, respectively). For all three causal structures $r_{X1Y} = r_{YX2} = 2/3$.

Eqs. (1)–(4) are standardized equations with variables centered on means of 50, representing different types of inferences for the chain $[X_1 \rightarrow Y \rightarrow X_2]$ and common cause $[X_1 \leftarrow Y \rightarrow X_2]$. Eq. (1) shows how to calculate the expected value of X_i given Y. Eq. (2) shows how to calculate the expected value of X_i given Y and X_j; X_j falls out of the equation reflecting the Markov Assumption that X_i is independent of X_j once Y is known. Eq. (3) is the transitive inference, and Eq. (4) is the inference to the "middle" variable. E stands for expected value.

$$\text{One-Link Inference.} \quad E(X_i|Y = y) = 50 + \left(\frac{2}{3}\right)(y - 50) \tag{1}$$

$$\text{Markov Assumption.} \quad E(X_i|Y = y, X_j = x_j) = E(X_i|Y = y) = 50 + \left(\frac{2}{3}\right)(y - 50) \tag{2}$$

$$\text{Transitive Inference.} \quad E(X_i|X_j = x_j) = 50 + \left(\frac{4}{9}\right)(x_j - 50) \tag{3}$$

$$\text{Middle Inference.} \quad E(Y|X_i = x_i, X_j = x_j) = 50 + \left(\frac{6}{13}\right)(x_i - 50) + \left(\frac{6}{13}\right)(x_j - 50) \tag{4}$$

For the common effect structure $[X_1 \rightarrow Y \leftarrow X_2]$, since $r_{X1Y} = r_{YX2} = 2/3$, the one-link inferences are the same as in Eq. (1). However, the other inferences are different. Since X_1 and X_2 are unconditionally independent, $r_{X1X2} = 0$, the best estimate of X_i given X_j is simply its mean, 50, Eq. (5). This is the inference relevant to the Markov Assumption for the common effect. Eq. (6) is the middle inference on the common effect structure, which comes straight from the parameters $r_{X1Y} = r_{YX2} = 2/3$. Eq. (7) is the explaining away inference; the negative coefficient on X_j is the key effect. Eq. (7) can be derived from the parameters using probability calculus.

$$\text{Markov Assumption Inference.} \quad E(X_i|X_j = x_j) = 50 \tag{5}$$

$$\text{Middle Inference.} \quad E(Y|X_i = x_i, X_j = x_j) = 50 + \left(\frac{2}{3}\right)(x_i - 50) + \left(\frac{2}{3}\right)(x_j - 50) \tag{6}$$

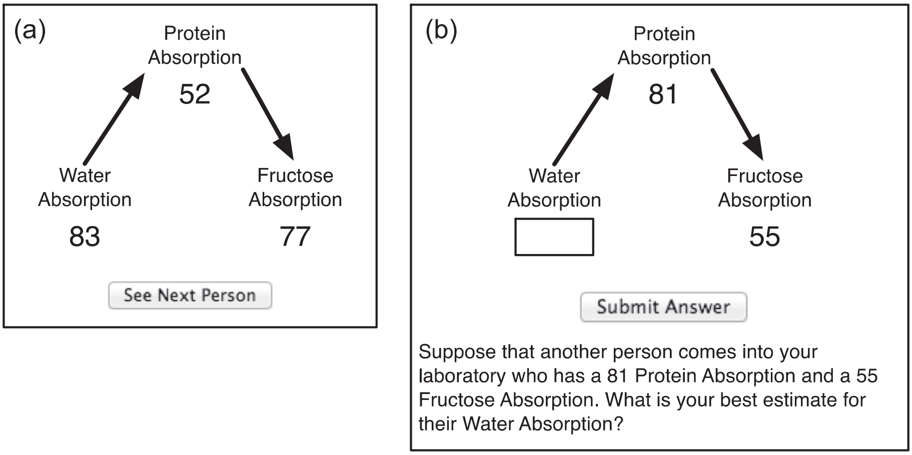
$$\text{Explaining Away.} \quad E(X_i|Y = y, X_j = x_j) = 50 + \left(\frac{6}{5}\right)(y - 50) - \left(\frac{4}{5}\right)(x_j - 50) \tag{7}$$

For the learning data all participants saw the same 32 trials (in a randomized order); this set of trials was constrained so that the multivariate distribution nearly perfectly matched the parameters above.

During the test phase the values of the known variables were chosen randomly from a multivariate normal distribution with the same parameters as in the learning phase, but each participant received a unique set of questions in order to sample broadly from the multivariate distribution.

3.2.3. Procedures

The procedures were the same as in Experiment 1, except for the following changes. In the learning phase participants saw numbers representing the magnitudes of the three variables (Fig. 7a). In the judgment phase (Fig. 7b) participants were asked to infer each numerical variable given information about one or both of the other variables. Participants made 36 inferences on each structure; see Table 8. Participants typed their responses as numbers from 0 to 100. (Due to a computer error 50 of the 7200 responses were not recorded.) At the end of the study, participants were paid a bonus for the number of questions that they answered within 3 points of the correct response.



Suppose that another person comes into your laboratory who has a 81 Protein Absorption and a 55 Fructose Absorption. What is your best estimate for their Water Absorption?

Fig. 7. Example screenshots in Experiment 2. Note: Panel (A) shows an example of one trial in the learning phase. Panel (B) shows an example of how participants made an inference of the type E(water|fructose = 55, protein = 81).

Table 8
Inference questions in Experiment 2.

Inference type	Specific inferences	Normative regression weight	Number of questions
Chain [$X_1 \rightarrow Y \rightarrow X_2$] and common cause [$X_1 \leftarrow Y \rightarrow X_2$]			
Markov Assumption	$E(X_i \vert Y = y, X_j = x_j)$	$Y = 2/3, X = 0$	12
Transitive	$E(X_i \vert X_j = x_j)$	$X = 4/9$	6
Middle	$E(Y \vert X_i = x_i, X_j = x_j)$	$X = 6/13$	6
One-link	$E(X_i \vert Y = y)$, $E(Y \vert X_i = x_i)$	$X = Y = 2/3$	12
Common effect [$X_1 \rightarrow Y \leftarrow X_2$]			
Explaining away	$E(X_i \vert Y = y, X_j = x_j)$	$Y = 6/5, X = -4/5$	12
Markov Assumption	$E(X_i \vert X_j = x_j)$	$X = 0$	6
Middle	$E(Y \vert X_i = x_i, X_j = x_j)$	$X = 2/3$	6
One-link	$E(X_i \vert Y = y)$, $E(Y \vert X_i = x_i)$	$X = Y = 2/3$	12

Note: The variables on the right hand side of the vertical bar were given as specific numbers. For example, one specific Markov Assumption question could have been $E(X_1 | y = 81, x_2 = 55)$.

3.3. Results

3.3.1. Success of the standard model

Agreeing with past reports, the standard linear model again accounts for a substantial portion of the variance in human judgments in Experiment 2, $r^2 = 0.36$ (Fig. 2). The following sections focus on the ways that the judgments deviate from the model.

3.3.2. Markov Assumption

For the chain and common cause, the Markov Assumption was tested by running regressions to test whether X_j had any effect on the inference of X_i when the state of Y is known $E(X_i|Y = y, X_j = x_j)$. Normatively every 1 point increase in Y should produce a .66 increase in X_i, and X_j should have no effect on X_i.

Specifically, for the inference $E(X_i|Y = y, X_j = x_j)$, the regression in Eq. (8) was fit with by-subject random effects on the intercept and random effects on the slopes to account for the repeated measures within subjects. The index k represents the 12 inferences clustered within the $l = 1$–50 subjects. α_l is the subject-specific intercept, and β_{Yl} and β_{Xjl} are the subject specific regression weights on Y and X_j. α_l, β_{Yl} and β_{Xjl} are all modeled as normally distributed random effects.

Table 9
95% CIs for regression weights testing the Markov Assumption in Experiment 2.

Regression weight	Normative	$X_1 \rightarrow Y \rightarrow X_2$	$X_1 \leftarrow Y \rightarrow X_2$	$X_1 \rightarrow Y \leftarrow X_2$
Y	0.66	[0.48, 0.75]	[0.32, 0.61]	–
X	0.00	[−0.06, 0.20]	[0.06, 0.31]	[.23, .48]

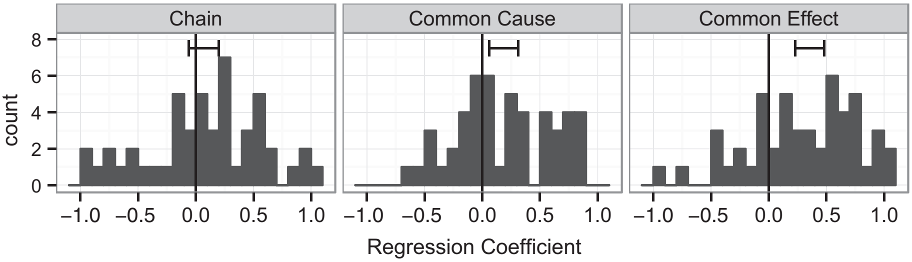


Fig. 8. Experiment 2 Markov Violations. Note: 5 regression coefficients outside the range [−1, 1] were plotted either as −1 or 1. Vertical lines are the normative answers. Horizontal lines are 95% CIs from Table 10.

$$x_{ikl} = \alpha_l + \beta_{Yl} Y_{kl} + \beta_{X_jl} X_{jkl} \tag{8}$$

Table 9 gives 95% confidence intervals for these regression weights. For the chain, the regression weights for X_j were not significantly different from 0, though X_j was trending in the positive direction. For the common cause, the regression weight for X_j was higher than 0, implying a violation of the Markov Assumption.

For the common effect, we analyzed inferences of $E(X_i|X_j = x_j)$; normatively X_i and X_j should be unconditionally independent. A regression with by-subject random effects on the intercept and random effects for the slope for X_j found that the participants did use X_j to predict X_i (see Table 9).

Fig. 8 shows a histogram of the distribution of regression weights on X_j. These regression weights and those in all the following figures are not the β_{Xjl} regression weights from Eq. (8). Instead they were calculated by running separate regressions for each participant to show the best-fitting regression weights for each participant; "no pooling" in the terminology of Gelman and Hill (2006). Fig. 8 shows considerable variation across participants, with many values far from the normative zero-impact coefficient and most in the positive direction.

Examining the distributions of regression weights for the chain in Fig. 8 reveals a potential reason why there were no significant violations of the Markov Assumption for the chain when analyzing all participants together. Within the chain condition, even though most participants have regression weights greater than 0 for the screened-off variable, which represents the positive Markov violation effect, there were 8 participants with regression weights less than −.50. It is possible that these subjects reflect a sub-group of participants who actually exhibit negative violations of the Markov Assumption instead of positive. For the chain [$X_1 \rightarrow Y \rightarrow X_2$], if $Y = 50$, a negative effect on X_2 means that the lower X_2 is, the higher the inference of X_1, and vice versa. So if $y = 50$ and $x_2 = 30$, a negative influence of X_2 implies a fairly high response for X_1 such as 70. This inference habit would create patterns of increasing or decreasing trends across X_1, Y, and X_2 (e.g., [$x_1 = 30 \rightarrow y = 50 \rightarrow x_2 = 70$]; [$x_1 = 75 \rightarrow \ y = 65 \rightarrow x_2 = 55$]). Future research could investigate whether there is a group of participants who consistently show this anomalous pattern of Markov Violations. If so, it is possible that the reason that the violation of the Markov Assumption was not significantly positive for the chain was because the two types of violations are canceling each other out.

3.3.3. Strength of inferences

Table 10 presents the 95% confidence intervals of regression weights for one-link, transitive, and middle inferences (compare to Eqs. (1), (3), (4), (6), and (7)). All of these regressions used

Table 10
95% CIs for regression weights for strength of judgments in Experiment 2.

Inference	Norm.	$X_1 \rightarrow Y \rightarrow X_2$	$X_1 \leftarrow Y \rightarrow X_2$	Norm.	$X_1 \rightarrow Y \leftarrow X_2$
$E(X_i \mid Y = y)$, $E(Y \mid X_i = x_i)$; one-link	0.66	[0.41, 0.61]	[0.54, 0.66]	0.66	[0.51, 0.69]
$E(X_i \mid X_j = x_j)$; transitive	0.44	[0.25, 0.47]	[0.51, 0.70]	[a]	[a]
$E(Y \mid X_i = x_i, X_j = x_j)$; middle	0.46	X_i: [0.25, 0.43] X_j: [0.30, 0.50]	X_i: [0.27, 0.49] X_j: [0.23, 0.45]	0.66	X_i: [0.42, 0.54] X_j: [0.40, 0.50]

[a] For the common effect, this inference is pertinent to the Markov Assumption and is discussed in that section.

by-subject random effects for the intercept and random effects for the slope(s) of the known variables analogous to Eq. (8). As always, there is considerable across-participant variation. The average one-link inferences tended to be too weak; some were significantly too weak and others were just trending. The transitive inferences on a causal chain were not different from normative, but for the common cause they were actually too strong.

Fig. 9 presents the "no-pooling" version of the analysis of strengths; the regression weight was calculated separately for each subject. Fig. 9 also includes the 95% confidence interval of the random-effects regression. The main impressions of Fig. 9 are that there is considerable variance in the regression weights of individual subjects and, with the one possible exception for the transitive inference on the common cause structure, the individual analysis agrees with the overall analysis such that the distributions of the regression weights are fairly symmetric. As will be seen in the results on the explaining away inference, this symmetry is not always found.

We were concerned that participants might be using a restricted portion of the response scale, particularly for the one-link, transitive, and middle inferences, since the results above suggest some conservatism. Participants were allowed to make inferences on the scale 0–100, where the normative answers were rarely outside the range 20–80. For the one-link, transitive, and middle inferences combined, the distribution of the normative answers was $M = 49.5$, $SD = 11.9$ for the chain, and $M = 50.1$, $SD = 11.9$ for the common cause. However, participants' responses were distributed with $M = 47.8$, $SD = 17.7$ for the chain and $M = 49.4$, $SD = 17.7$ for the common cause; the standard deviations were too large. Furthermore, the standard deviation of each individual participant's judgments was always higher than the standard deviation of the normative answers for the given participant except for three cases out of 100.[9] This pattern is remarkably consistent; our participants used more of the scale than is warranted. These results imply that the conservative tendency is not due to reluctance to use the entire scale.[10]

In sum, the analysis of the strength of inferences is inconsistent with the representativeness heuristic, that people would respond with a magnitude equally as extreme as the provided cue (Kahneman & Tversky, 1973). A strict interpretation of representativeness implies regression weights of 1. Though there were some subjects with regression weights of 1 or higher, 1 was considerably outside all of the 95% confidence intervals on the overall regression weight. For the most part, participants' inferences were appropriately regressive or a bit overly regressive (conservative), more consistent with the findings of Lichtenstein et al. (1975). Similar to the study of Lichtenstein et al., and in contrast to the study by Kahneman and Tversky, our study did involve trial-by-trial learning from experience.

[9] There was one participant who gave 50 as the response to every question for the chain, one who gave 55 for every response to the common cause, and another participant whose judgments had a lower standard deviation than the normative judgments for the chain.

[10] It might initially seem that this finding of higher standard deviations in participants' judgments than the normative responses is inconsistent with the finding of conservative inferences. However, this is possible statistically. Consider a variable $A \sim N(M = 0, SD = 1)$, and let $B = .5A + \varepsilon$, where $\varepsilon \sim N(M = 0, SD = 1)$. Think of A as the normative answer, and B as a participant's response; B is only half as strong as it should be, and there is also error in B. In this case the standard deviation of B, sqrt(5/4) = 1.12, is greater than the standard deviation of A even though B is too conservative. A linear regression using A to predict B will reveal a regression weight of .5 on average, correctly recovering the degree of conservatism. This is essentially how the analyses above work, except that instead of using the normative response as the predictor of participants' judgments, we used the states of the one or two variables that were known to predict participants' judgments. Also, note that having high error ε would reduce the precision of the estimate of the slope, but does not affect the point estimate of the slope. Restated, high error of this form would not appear as a conservative (or anticonservative) bias.

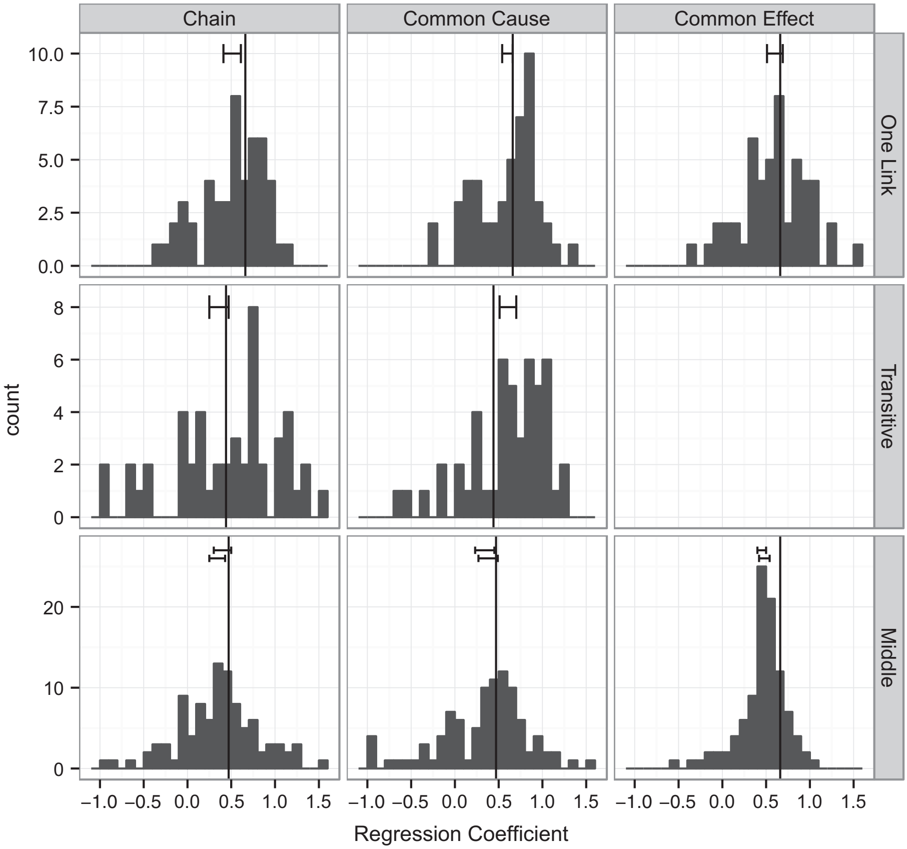


Fig. 9. Strength of regression weights in Experiment 2. Note: 15 regression coefficients outside the bounds of [−1, 1.5] were plotted as −1 or 1.5. Vertical lines represent the normative point-estimate model answers. Horizontal lines are 95% CIs. The transitive common effect graph is included in the Markov Assumption section.

3.3.4. Explaining away

Explaining away was analyzed by examining the $E(X_i|Y, X_j)$ inference on the common effect structure using Eq. (8). The 95% confidence interval for $Y = [0.87, 1.20]$, and the confidence interval for $X_j = [-0.42, -0.14]$. The fact that the confidence interval for X_j is entirely less than zero implies that there was a significant explaining away effect. However, the lower end of the confidence interval, −0.42, is only about half as strong as the normative value of −0.80 (i.e., there seems to be a grasp of the basic principle of explaining away, but inferences following that principle are much less strong than they ought to be normatively).

Fig. 10 shows a histogram of coefficients for X_j when separate regressions are run for each participant. There is considerable variance and skew, but the distribution supports a more optimistic view of participants' adherence to normative explaining away principle, than the overall regression with random effects on the slope. That is, the mode of the distribution is around −0.50 and the median was −.44. Still, only 5 out of the 50 participants had regression weights less than the normative value of −0.80, though the coefficients should be centered on −0.80 if participants were following the normative principle.

4. General discussion

A literature review and two experiments found that from a global perspective, the standard rational Causal Bayesian Network model of causal reasoning accounts for a substantial portion of the variance in human judgments (Fig. 2). However, a finer-grained analysis of individual judgments identified three consistent violations of the standard model in addition to some other smaller violations: (1) violations of the Markov Assumption, (2) some inferences are too weak, whereas others are too strong, and (3) insufficient explaining away. In Section 4, we first review the evidence for these conclusions. Then we discuss possible theories and models to explain the findings. Lastly, we discuss the significance of the behavioral findings.

4.1. Summary of results

The findings are most clear cut for the Markov Assumption and explaining away, and more subtle for the strengths of inferences.

4.1.1. Markov Assumption

Markov Violations[11] were observed at exactly the same rates for binary and numerical events, indexed by effects significant (at the $p < .05$ level) in the "positive" direction (meaning violations). For the binary variables in Experiments 1a and 1b, four out of six tests showed significantly positive violations, and the other two were in the positive direction. For the numerical variables in Experiment 2, two of the three tests showed significantly positive violations, and the third was in the positive direction.

4.1.2. Explaining away

Explaining away was insufficient across both binary and numerical events; in Experiment 1a the effect was somewhat in the wrong direction, whereas in Experiments 1b and 2 the effect was in the right direction but was too small.

4.1.3. Strength of inferences and spikes for binary variables

Based on prior results, we predicted that inferences would be too weak with binary variables, but would be too strong for numerical variables because participants might anchor on a specific cue value and fail to adjust (regress) sufficiently towards the mean. However, in hindsight, this hypothesis was overly simplistic; the responses depend on the particular type of inference. We first review the inferences on binary variables.

The inferences on binary variables that were too weak and had spikes at .5 occurred for two reasons. (Here the 'too weak' results and the spikes at .5 are essentially the same finding.) First, they tended to be weak when the two known variables conflicted, such as the 'low' Markov Violation inference $P(x_i = 1|y = 1, x_j = 0)$ on the chain and common cause (Table 4, Fig. 3). Second, they tended to be too weak when the state of one variable was unknown. One example is the transitive inference $P(x_i = 1|x_j = 1)$. The transitive inferences were only too weak when the normative inference was not too close to .5; for Experiment 1a the normative inference was .625, and the judgments were not significantly different from normative, but for Experiment 1b when the normative inference was .78 they were significantly too weak (Table 6, Fig. 4). Another example of weakness when one variable was unknown is the inference of $P(x_i = 1|y = 1)$ for the common effect structure (Table 7, Fig. 6). Again, this inference was only too weak when the normative inference was not too close to .5; this time for Experiment 1a (normative = .71) but not for 1b (normative = .60).

In contrast, the inferences that tended to be too strong were those for which the two cues were consistent, including (1) the Markov violation inference $P(x_i = 1|y = 1, x_j = 1)$ on the chain and common cause for Experiment 1a (Table 4, Fig. 3), (2) the middle variable inference $P(y = 1|x_i = 1, x_j = 1)$ on the

[11] One might argue that Markov violations were encouraged by asking participants to make many inferences in a brief space of time. However, participants were paid in proportion to the time that they took on the experiment, and there was no explicit time pressure. In addition, Rehder (2014) found that increasing time pressure did not affect the rate at which the Markov Assumption was violated.

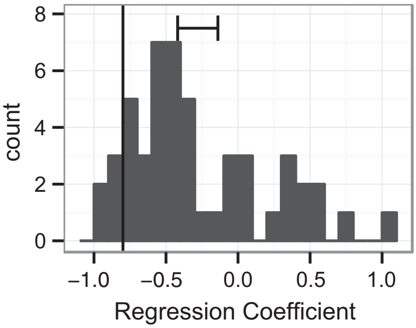


Fig. 10. Explaining away in Experiment 2. Note: One coefficient >1 was plotted as 1. The vertical line is the normative answer, −0.8. Horizontal bar is 95% CI of regression coefficient.

common effect $X_1 \rightarrow Y \leftarrow X_2$ structure (Table 6, Fig. 4), and (3) the explaining away inference on the common effect $X_1 \rightarrow Y \leftarrow X_2$ when both cues were present $P(x_i = 1|y = 1, x_j = 1)$ (Table 7, Fig. 6). One benefit of this explanation is that the violations of the Markov Assumption are incorporated into a larger pattern of strength versus weakness due to consistent versus inconsistent cues, rather than being treated as a unique phenomenon.

There are two exceptions to the pattern of too strong inferences on binary variables: (1) the middle variable inference $P(y = 1|x_i = 1, x_j = 1)$ on the chain and common cause, which were too weak (Table 6, Fig. 5), and (2) the Markov Violation inference $P(x_i = 1|y = 1, x_j = 1)$ for the chain and common cause in Experiment 1b (Table 4, Fig. 3), which were not significantly different than normative. For these two exceptions, the normative inferences were .875 or greater.

Overall, these patterns of too strong versus too weak inferences can be described as too weak when the state of one cue is unknown (‘ambiguity aversion’) or when the two cues conflict (‘conflict aversion’), and too strong when the two cues are consistent (the opposite of ‘conflict aversion’). This pattern is moderated by whether the normative inference is close to .5 or is extreme (close to 1 or 0). When the normative inference was close to .5, it was harder to see the too weak inferences (because the normative inferences were already weak, so there is essentially a floor effect) but easier to see the too strong inferences (because there is lots of room higher in the probability scale). In contrast, when the normative inference was close to 1, it was easier to see the too weak inferences and harder to see the too strong inferences.

Once the moderating effect of scale use is considered, the findings of which inferences were too strong versus too weak (and spikes at .5) line up closely with the heuristics of ambiguity aversion and conflict aversion. Below we include these simple heuristics as part of a longer list of possible explanations for the findings.

4.1.4. Strength of inferences on numerical variables

One important finding is that when looking at the results in Tables 9 and 10, the regression weights that should normatively be zero tend to be weaker than those than should normatively be .44 or .46, which tend to be weaker than those that should normatively be .66. Though this effect was not tested statistically, it implies a degree of sensitivity to the normative strength (although, many of the confidence intervals are overlapping, implying that this effect is not robust).

For the binary variables, inferences tended to be too weak when one of the cues was unknown. For this reason, we could expect the one-link and transitive inferences (Table 10) to also be too weak for the numerical variables. Out of the five transitive and one-link variables, one of them was significantly too weak, one was significantly too strong, and the other three were in the direction of being too weak but were not significantly different from normative. It is possible that using learning data with stronger correlations between causes and effects would result in inferences that are too weak, but in the current experiment the effect of a missing cue is not reliable.

For binary variables, inferences tended to be too strong when the cues were consistent, and too weak when they were inconsistent. However, this explanation does not map easily to numerical variables because it would be very unlikely for two numerical cues on the scale 1–100 to be exactly the same. Applying these explanations to numerical variables would require specifying how close is close enough to be ‘consistent.’ For this reason, we have not gone farther in trying to map this heuristic onto continuous variables.

Finally, we consider numerical inferences on the middle variable. Out of the six inferences in Table 10, four were significantly too weak, and the other two were in the direction of being too weak but were not significantly different from normative. It is possible that one reason for the weakness is that learners may often perceive the two cues to be inconsistent, and hedge; however, this is just a hypothesis, and there may be other explanations for the weakness when inferring a middle variable.

4.2. Explanations for the findings

A considerable challenge for understanding the reasons for the current findings is that there are so many possible explanations. In Sections 4.2.1–4.2.5 and Table 11, we attempt to comprehensively explain how these existing theories (and some new theories) could account for the current results. The reason for being comprehensive is that in previous research, models have often been assessed only insofar as they help explain one particular type of judgment. Here we try to assess all the important previously-proposed theories on many different inference types.

The previous literature has focused on explanations for the violations of the Markov Assumption (Models 2–6). We discuss these theories not just for the Markov Assumption, but also for the predictions they make for other inferences. We then introduce three simple heuristic models (Models 7–9); these sorts of heuristics are often used to explain performance in multiple cue judgment tasks but have typically not been used for causal judgments. Lastly, we discuss three additional unrelated models (10–12), one of which we invented (Model 11). This analysis focuses on binary variables, as many of the proposals have only considered binary events, and plausible accounts for numerical variables would require substantial changes. The standard Bayesian model (Theory 1) cannot account for any of the findings in Table 11.

4.2.1. Models 2–6: alternative causal structures

The most prominent explanation for violations of the Markov Assumption has been to presume that participants do not fully accept the causal structure presented by the experimenter, and then participants add additional events (nodes) and causal relations to the mental network that guides their reasoning (Mayrhofer & Waldmann, 2015; Meder, Mayrhofer, & Waldmann, 2014; Park & Sloman, 2013; Rehder, 2014; Rehder & Burnett, 2005). Table 11 includes four possibilities, Theories 2–5, which are specified in Fig. 11, adapted from Rehder (2014). Each of these theories adds one or more nodes to the causal structure in a relatively analogous way across the structures. The difference between the specific versus general disabler accounts is whether the disabler acts on the background causes in addition to all the observed causes, or only on the observed causes. (Code to simulate these models to generate the predictions in Table 11 can be obtained by contacting the authors.)

Out of these four modifications to the standard Bayesian Networks account, the Specific Shared Disabler was best supported by one previous study (Park & Sloman, 2013). This study found that there were only violations of the Markov Assumption when the middle variable was present, not when it was absent, which is a unique feature of the specific shared disabler account. (Another study found some support for this shared disabler account Rehder, 2014, p. 80.) However, we found that the violations were roughly similar when the middle variable was present or absent. The results broken down by these two conditions can be found in Appendix A. Therefore, our results do not support the specific shared disabler account.

Across Theories 2–5, Table 11 shows that the shared generative cause approach explains the largest number of the findings, though it still cannot explain all of them. Furthermore, none of these models captures the spikes at exactly .50 when the two cues conflict like $P(x_1 = 1|y = 1, x_2 = 0)$ on the chain and common cause. (This finding is one of the insights from analyzing the distributions of the judgments, which were not reported in previous studies.)

Table 11
Which theories can explain which empirical phenomena for Experiments 1a and 1b with binary variables.

Theory	Markov violations on chain and common cause (Fig. 3, Table 4) $X_1 \leftarrow Y \rightarrow X_2$, $X_1 \rightarrow Y \rightarrow X_2$ $P(x_i = 1 \vert y = 1, x_j = 1)$ versus $P(x_i = 1 \vert y = 1, x_j = 0)$	Markov violations on common effect (Fig. 3, Table 5) $X_1 \rightarrow Y \leftarrow X_2$ $P(x_i = 1 \vert x_j = 1)$ versus $P(x_i = 1 \vert x_j = 0)$	Weak transitive inferences (Fig. 4, Table 6) $X_1 \leftarrow Y \rightarrow X_2$, $X_1 \rightarrow Y \rightarrow X_2$ $P(x_i = 1 \vert x_j = 1)$ versus $P(x_i = 1 \vert x_j = 0)$	Overly strong middle inferences for common effect (Fig. 5, Table 6) $X_1 \rightarrow Y \leftarrow X_2$ $P(y = 1 \vert x_i = 1, x_j = 1)$	Weak explaining away on common effect (Fig. 6, Table 7) $P(x_i = 1 \vert y = 1, x_j = 1)$ versus $P(x_i = 1 \vert y = 1)$ versus $P(x_i = 1 \vert y = 1, x_j = 0)$
Models 1–5: The standard Causal Bayesian Network (CBN) model with 5 modifications					
1. Standard CBN	No	No	No	No	No
2. Standard CBN plus **Specific Shared Disabler** (Fig. 11). A specific disabler only disables observed causes, not background causes	Partly. This account predicts Markov violations for the chain and common cause only when the middle variable is present. In the current studies Markov violations were also found when the middle variable was absent. Also does not predict spikes at .50	No	No. Predicts that $P(x_i = 1 \vert x_j = 0)$ would be even lower, farther away from .5	No. Predicts that $P(y = 1 \vert x_i = 1, x_j = 1)$ would be too low instead of too high	No. Does not have an influence on explaining away. In these studies, there was no background cause for the common effect Y, so there is no difference between the general versus specific disabler account for the common effect
3. Standard CBN plus **General Shared Disabler** (Fig. 11). A general disabler acts on all observed causes and background causes	Poorly. Accounts for violations in common cause. But for the chain, either does not predict any violation, or prediction is in the wrong direction. See Rehder (2014) for more details. Also does not predict spikes at .50	No	No. Predicts that $P(x_i = 1 \vert x_j = 0)$ would be even lower, farther away from .5	No. Predicts that $P(y = 1 \vert x_i = 1, x_j = 1)$ would be too low instead of too high	No. Does not have an influence on explaining away
4. Standard CBN plus **Shared Mediator** (Fig. 11)	Partly. Does not predict spikes at .50. Also, Markov violations for the chain would be very small (see Rehder (2014))	No	No. Predicts that $P(x_i = 1 \vert x_j = 1)$ would be even higher, farther away from .5	Potentially if the mediator is strongly generative. However, this account also predicts that $P(y = 1 \vert x_i = 1, x_j = 0)$ would be overestimated, but it was not	Somewhat. Adding a mediator essentially adds noise to Y, which reduces explaining away. But cannot explain why $P(x_i = 1 \vert y = 1, x_j = 1) > P(x_i = 1 \vert y = 1)$ in Experiment 1
5. Standard CBN plus **Shared Generative Cause** (Fig. 11)	Partly. Explains Markov Violations but does not predict spikes at .50	Yes	No. Predicts that $P(x_i = 1 \vert x_j = 1)$ would be even higher, farther away from .5	Yes	Somewhat. Adding a shared generative cause can reverse explaining away such that $P(x_i = 1 \vert y = 1, x_j = 1) > P(x_i = 1 \vert y = 1) > P(x_i = 1 \vert y = 1, x_j = 0)$. However, this model does not explain the pattern in Experiment 1: $P(x_i = 1 \vert y = 1, x_j = 1) > P(x_i = 1 \vert y = 1) < P(x_i = 1 \vert y = 1, x_j = 0)$

(*continued on next page*)

Table 11 (*continued*)

Theory	Markov violations on chain and common cause (Fig. 3, Table 4) $X_1 \leftarrow Y \rightarrow X_2$, $X_1 \rightarrow Y \rightarrow X_2$ $P(x_i = 1 \mid y = 1, x_j = 1)$ versus $P(x_i = 1 \mid y = 1, x_j = 0)$	Markov violations on common effect (Fig. 3, Table 5) $X_1 \rightarrow Y \leftarrow X_2$ $P(x_i = 1 \mid x_j = 1)$ versus $P(x_i = 1 \mid x_j = 0)$	Weak transitive inferences (Fig. 4, Table 6) $X_1 \leftarrow Y \rightarrow X_2$, $X_1 \rightarrow Y \rightarrow X_2$ $P(x_i = 1 \mid x_j = 1)$ versus $P(x_i = 1 \mid x_j = 0)$	Overly strong middle inferences for common effect (Fig. 5, Table 6) $X_1 \rightarrow Y \leftarrow X_2$ $P(y = 1 \mid x_i = 1, x_j = 1)$	Weak explaining away on common effect (Fig. 6, Table 7) $P(x_i = 1 \mid y = 1, x_j = 1)$ versus $P(x_i = 1 \mid y = 1)$ versus $P(x_i = 1 \mid y = 1, x_j = 0)$
6. Standard CBN with **Priors on Causal Strength** parameters	No. Changes to the parameters do not change the independencies implied by the structure	No. Changes to the parameters do not change the independencies implied by the structure	No. Priors can only explain overly weak or overly strong inferences, not both simultaneously. Also, the priors estimated from the previous literature (Lu et al., 2008) barely have any influence on the inferences given the moderate amounts of experience in these studies		
Models 7–9: Three heuristic models					
7. **Ambiguity Aversion**. When the state of one cue is unknown, judgments become uncertain (close to .5 and or spiked at .5)	NA. For these judgments both cues are known	Partly. In both judgments one cue is unknown, so it cannot explain the difference between the two in Experiment 1b. However, it can explain the spikes at .50 for Experiment 1b in which the normative answer is .25	Yes. In transitive judgments the middle variable is not known. These judgments tend to have peaks at .5	NA. For these judgments both cues are known	Partly. This could explain the peaks at .5 for the $P(x_i = 1 \mid y = 1)$ judgments
8. **Conflict Aversion**. When two cues conflict (1 and 0), judgments become uncertain (close to .5 and or spiked at .5)	Yes. When the two cues conflict there are spikes at .5	NA. Only one cue is known, so there cannot be any conflict	NA. Only one cue is known, so there cannot be any conflict	Yes. The concordance of the two cues (lack of conflict) is consistent with high judgments	Yes. This can help explain why some of the $P(x_i = 1 \mid y = 1, x_j = 0)$ judgments are lower than they should be. Additionally, the lack of conflict between cues could explain why the $P(x_i = 1 \mid y = 1, x_j = 1)$ judgments are so high
9. **Monotonicity Assumption**. In situations with positive causal relations, more present (absent) cues leads to higher (lower) judgments	Yes. This theory predicts that $P(x_i = 1 \mid y = 1, x_j = 1) > P(x_i = 1 \mid y = 1, x_j = 0)$. For $P(x_i = 1 \mid y = 1, x_j = 0)$, this theory predicts spikes at .5	Partly. This theory predicts that $P(x_i = 1 \mid x_j = 1) > P(x_i = 1 \mid x_j = 0)$ even though they should be equivalent. This theory does not explain why this effect occurred for Experiment 1b but not 1a	NA. This effect does not involve a comparison of two judgments; it is just about one judgment	Yes. The judgment of $P(y = 1 \mid x_i = 1, x_j = 1)$ is very high for the common effect, and looks similar to the chain and common cause	Partly. This theory predicts the exact opposite pattern compared to explaining away: $P(x_i = 1 \mid y = 1, x_j = 1) > P(x_i = 1 \mid y = 1) > P(x_i = 1 \mid y = 1, x_j = 0)$. Though this order was not found, this relationship could contribute to the failure to find explaining away

Models 10–12: Three unrelated models					
10. **Undirected Graphical Models** (aka Markov Networks) (see Koller and Friedman, 2009, pp. 139–142; Murphy, 2012, pp. 664–665; Rehder, 2014)	No. Chains and common cause networks can be perfectly represented by an undirected GM. This means that a $X_1 - Y - X_2$ structure upholds the Markov Assumption	Yes. UGMs cannot perfectly represent $X_1 \rightarrow Y \leftarrow X_2$. If it is represented as $X_1 - Y - X_2$, then it treats X_1 and X_2 as unconditionally positively correlated. Alternatively, if an additional $X_1 - X_2$ negative link is added to capture negative conditional dependence, X_1 and X_2 may be nearly unconditionally independent	No. UGMs can perfectly capture chains and common cause structures, so they do not explain why inferences would be weak	Complicated. This depends on whether the model is represented as $X_1 - Y - X_2$ or whether it includes an additional negative $X_1 - X_2$ link. It also depends on how the parameters are learned from the data, which is much more complicated than for UGMs	Yes. UGMs cannot perfectly represent $X_1 \rightarrow Y \leftarrow X_2$. If it is represented as $X_1 - Y - X_2$, then it treats X_1 and X_2 as conditionally independent given Y
11. **Beta Inference** directly from the data plus sampling from the posterior. Mean predictions are calculated for Experiment 1a	Partly. The skew in the distributions and unequal amounts of observations leads to different means for the two inferences $P(x_i = 1 \| y = 1, x_j = 1) = .71$ $P(x_i = 1 \| y = 1, x_j = 0) = .66$ However, it does not specifically predict spikes at .50	Yes, due to skew and unequal numbers of observations. Furthermore, this model also explains why there is a small violation for Experiment 1b but not 1a Experiment 1a: $P(x_i = 1 \| x_j = 1) = .5$ $P(x_i = 1 \| x_j = 0) = .5$ Experiment 1b: $P(x_i = 1 \| x_j = 1) = .26$ $P(x_i = 1 \| x_j = 0) = .25$	Partly. The skew in the posterior distribution predicts judgments lie closer to .5 $P(x_i = 1 \| x_j = 1) = .61$ $P(x_i = 1 \| x_j = 0) = .39$ It does not specifically predict spikes at .50	No. The judgments are considerably stronger than this model $P(y = 1 \| x_i = 1, x_j = 1) = .7$	A bit. The skew in the posterior distribution makes the explaining away weaker than the standard account, but still predicts a sizeable explaining away effect $P(x_i = 1 \| y = 1, x_j = 1) = .58$ $P(x_i = 1 \| y = 1) = .68$ $P(x_i = 1 \| y = 1, x_j = 0) = .83$
12. **Nonlinear Transformation of Probability Judgments**	No. Any transformation would change both of the judgments to the same new value	No – see left	Partly. The standard S-shaped transformation brings judgments closer to the middle of the scale. This could account for weak transitive inferences, but then cannot account for the overly-strong middle inferences		Partly. Could explain why the three judgments are closer together than they should be. But cannot explain why $P(x_i = 1 \| y = 1, x_j = 1) > P(x_i = 1 \| y = 1)$ in Experiment 1

In sum, none of the standard accounts in which participants add beliefs to the causal structure can explain all of the findings; though some do better than others.

Another way to modify the standard CBN account is to add priors on the causal strengths (Theory 6). The idea is that when participants experience the contingency between a cause and effect, that they have priors about causal strength that bias their interpretation of causal strength (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Yeung & Griffiths, 2015). Priors on the strengths cannot explain the violations of the Markov Assumption because they do not change the independencies implied by the causal structure. However, they could potentially explain the weak transitive inferences, weak explaining away, or overly strong middle inference on the common effect. Still, there are a number of reasons why priors are a poor explanation for these effects.

First, the priors could explain either why inferences are too strong or why they are too weak, but they cannot explain why some of the inferences were too strong and others were too weak, both of which occur on the common effect structure. Second, the priors that have been empirically estimated in past research are uninformative enough that, given the data that participants experienced in the present studies, the priors could only have a very small influence. In particular, the causal strength estimates using the priors from Lu et al. (2008) produce values very close to the same estimates obtained with no priors (Cheng, 1997): .69 with prior versus .67 without for Experiment 1a, and .85 with versus .85 without the prior for Experiment 1b.

4.2.2. Models 7–9: simple judgment heuristics

A number of the findings can be explained by simple judgment heuristics, the types of heuristics that are often used to explain inference in multiple cue judgment tasks, though they have not traditionally been used to explain causal judgment.

Three of the documented effects involve situations in which the state of one of the cues is unknown. For example, the state of Y is unknown when making the judgment $P(x_1 = 1|x_2 = 1)$. In such instances we often found spikes of judgments at .50, suggesting that participants resort to the middle of the scale when they feel ignorant. This pattern of reasoning could be viewed as a type of ambiguity aversion (Theory 9) (Camerer & Weber, 1992).

Another situation in which participants often answer with .50 is in cases when two cues conflict with each other such as $P(x_1 = 1|y = 1, x_2 = 0)$. We called this pattern conflict aversion (Theory 10). The converse of this phenomenon is that when inferring the middle variable on the common effect and both causes were present, the judgments were too strong. We cannot find any citations to related phenomena in judgments directly comparable to those required in our causal reasoning tasks. But, there is a history of interpreting choice strategies for options like consumer goods, as motivated to minimize the need to deal with conflicting attributes of single options (e.g., the difficult trade-off between economy and safety or durability in a product like a car or a cellphone; e.g., Bettman, Luce, & Payne, 1998).

A third simple judgment heuristic that fits well with a number of the findings is that judgments tended to be monotonically related to the number of cues that are present minus the number of cues that are absent, which we call the monotonicity assumption (Theory 11). This heuristic predicts, for example, that $P(x_1 = 1|y = 1, x_2 = 1) > P(x_1 = 1|y = 1) > P(x_1 = 1|y = 1, x_2 = 0) > P(x_1 = 1|y = 0) > P(x_1 = 1|y = 0, x_2 = 0)$. In these studies only generative causal relations were studied, but this heuristic could be extended in cases when there are inhibitory causal relations or combinations of generative and inhibitory relations. A monotonicity assumption explains many of the same effects as ignorance aversion and conflict aversion combined. Gigerenzer and his colleagues, among others, have proposed similar "tallying" rules as short-cut heuristics for many judgments (Gigerenzer & Gaissmaier, 2011).

Overall, these simple heuristics can explain a number of the findings. Simple heuristics are a staple in the field of judgment and decision making research, but are rarely discussed in research on causal reasoning. The current findings suggest that it may be worthwhile to consider the role of some simple heuristics in causal reasoning.

Models 10–12 are each theoretically unique from all the other models, so they are discussed separately.

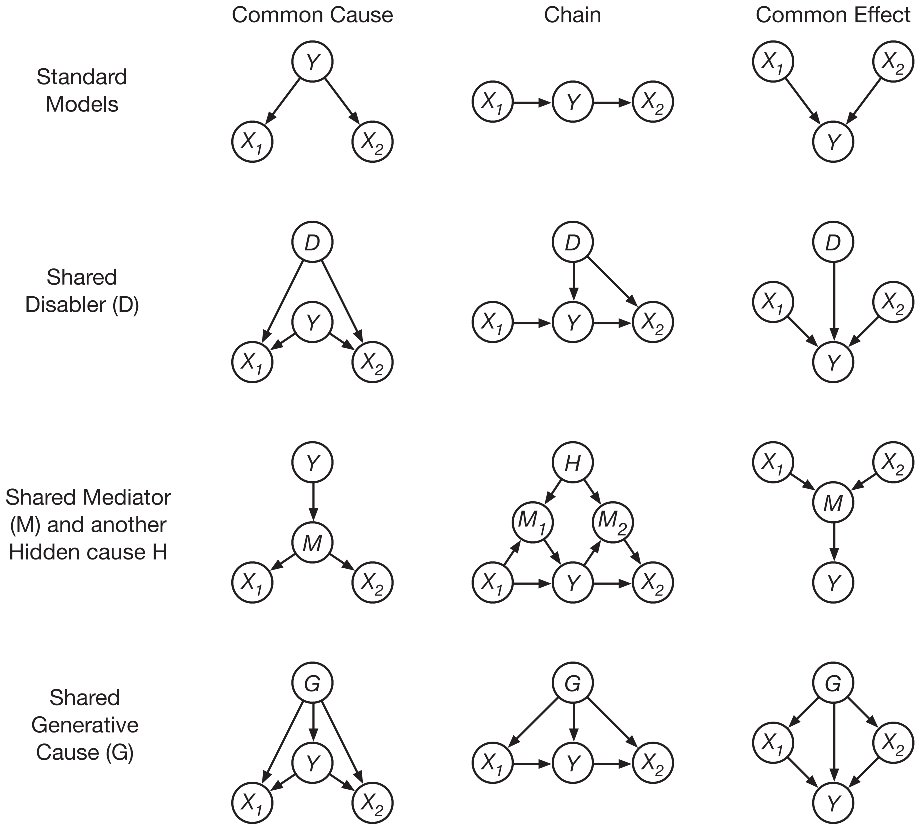


Fig. 11. Alternative Bayes nets representations. Adapted from Rehder (2014) Fig. 5.

4.2.3. Model 10: undirected graphical models

Rehder (2014) proposed that instead of reasoning with a directed graphical model, that some participants reason associatively. He modeled associative processes as if participants treat the network as an undirected graphical model (UGM; sometimes called a Markov random field; Koller & Friedman, 2009, chap. 4, especially pp. 139–142; Murphy, 2012, chap. 19, especially pp. 664–665). The multivariate probability distribution and all the dependencies and independencies inherent in common cause [$X_1 \leftarrow Y \rightarrow X_2$] and chain [$X_1 \rightarrow Y \rightarrow X_2$] networks can be perfectly represented as a [$X_1 - Y - X_2$] UGM because they are "chordal". This means that using an UGM representation does not explain the violations of the Markov Assumption, nor the weak inferences for the chain and common cause.

Furthermore, no UGM can perfectly capture the multivariate probability distribution and all the dependencies and independences for a common effect structure [$X_1 \rightarrow Y \leftarrow X_2$] (see citations above). This structure could be approximately represented as [$X_1 - Y - X_2$], either with or without with an additional [$X_1 - X_2$] link. If the common effect structure were to be represented as [$X_1 - Y - X_2$], then the explaining away effect would disappear. As for the middle inference on the common effect, we cannot make an unequivocal prediction because there are at least two possible UGM representations and because learning the parameters of an UGM from data is complicated and would require additional theoretical commitments (see citations above).

4.2.4. Theory 11: Beta inference

We introduce an original model of causal inference that could be used when reasoning about binary variables experienced from data. The motivation for the Beta Inference Model is that when the causal structure is supplemented with learning data, it is possible to make judgments directly from the data itself. For example, when making the judgment $P(x_1 = 1|x_2 = 1)$, a learner could simply divide the number of trials in which $x_1 = 1$ and $x_2 = 1$ by the number of trials in which $x_2 = 1$. These judgments would be exactly the same as the more typical version of the CBN theory in which the learner infers parameters from the learning data, and then uses the parameters to make judgments.

The Beta Inference Model goes one step farther, and instead of inferring a point estimate of the judgment, it infers a posterior distribution for the judgment directly from the data. For example, the inference $P(x_1 = 1|x_2 = 1)$ can be conceived as a problem of figuring out the proportion of times that $x_1 = 1$ (“success”) versus $x_1 = 0$ (“failure”) within the set of cases in which $x_2 = 1$. Given α successes and β failures, the posterior distribution of the proportion $\alpha/(\alpha + \beta)$ is defined by the Beta($\alpha + 1$, $\beta + 1$) distribution (Kruschke, 2011). (See Neapolitan, 2004, chap. 6, for a tutorial on using the Beta distribution for learning parameters on causal networks.) In Experiment 1a, which had 10 trials in which $x_1 = 1$ and $x_2 = 1$, and 6 trials in which $x_1 = 0$ and $x_2 = 1$, the posterior distribution for the inference $P(x_1 = 1|x_2 = 1)$ is Beta(11, 7). The model assumes that subjects sample from the posterior distribution when making the inference.

This approach can be used to calculate the posterior distribution of any inference from Table 2. For example, the posterior distribution for the inference $P(x_1 = 1|y = 1, x_2 = 0)$ uses the parameters $\alpha = N(x_1 = 1, y = 1, x_2 = 0)$ and $\beta = N(x_1 = 0, y = 1, x_2 = 0)$, where N means the number of trials.

One important feature of many of these Beta distributions is that they are skewed. The mode of a Beta($\alpha + 1$, $\beta + 1$) distribution is equal to the proportion $\alpha/(\alpha + \beta)$; this means that the peak of the distributions predicted by this model is always equivalent to the peak of the standard point estimate Bayesian model. However, the mean of a Beta($\alpha + 1$, $\beta + 1$) distribution is $(\alpha + 1)/(\alpha + \beta + 2)$; and the distributions are skewed such that the mean is always closer to .50 than the mode.

If participants make judgments by sampling from the posterior of the Beta($\alpha + 1$, $\beta + 1$) distribution, the skew of the distribution explains the weak transitive inferences, and explains some degree of weakness for explaining away. Additionally, this same account also explains violations of the Markov Assumption. For example, when inferring $P(x_1 = 1|y = 1, x_2 = 1)$ versus $P(x_1 = 1|y = 1, x_2 = 0)$, there are a larger number of trials in which $y = 1$ and $x_2 = 1$ than when $y = 1$ and $x_2 = 0$. This means that even though the modes of the distributions are the same (.75 for Experiment 1a), the distribution for $P(x_1 = 1|y = 1, x_2 = 0)$ is more skewed and thus its mean (.66) is closer to .5 than the mean of the distribution for $P(x_1 = 1|y = 1, x_2 = 1) = .71$. Graphs of the predictions of the Beta Inference Model compared to subjects’ judgments, as well as a more detailed analysis of the successes and failures of the model are presented in Appendix B.

In summary, though this new Beta model account cannot explain all the findings, it can explain a number of the findings.

4.2.5. Theory 12: nonlinear transformation of probability judgments

The last theory is that decision makers do not treat probabilities linearly. The standard implementation is the *S*-shaped decision weight function from prospect theory, in which low probabilities are under-estimated and high probabilities are over-estimated (Gonzalez & Wu, 1999; Kahneman & Tversky, 1979). A nonlinear probability scale cannot explain the violations of the Markov Assumption, because both probabilities would be transformed equally. It partially explains the weak transitive inferences and weak explaining away, because extreme probabilities are brought closer to the middle of the scale; however, it cannot explain the overly-strong middle judgment on the common effect, nor can it explain the reversal of probabilities in explaining away in Experiment 1.

4.2.6. Combinations of theories to explain the findings

We conclude that there is no single theory or explanatory principle yet proposed that can account for all the findings. On the other hand, there are multiple combinations of models that each can account for most of the results. For example, the monotonicity assumption plus a nonlinear transformation of probability judgments can do quite well. Alternatively, the Beta Inference Model plus conflict aversion, plus nonlinear transformation of probability judgments does quite well. There are likely numerous other combinations that can explain most of the results. This is a hard position for the field to be in because it makes it difficult to come to a satisfactory explanation for how people reason about causal structures.

It is possible that future work will be able to quantitatively compare all the combinations of these models. For example, Rehder (2014) performed a test with four theories and concluded that all four were involved in explaining the results. However, his search considered a more limited number of phenomena, and the judgments were qualitative rather than quantitative.

When considering the 12 theories in Table 11 (as well as modifications to these theories), the number of possible combinations is so large that we doubt that such a search could yield a strong conclusion that one particular combination really is the most successful. For this reason, we have decided not to attempt a quantitative model comparison search at the present time. We suspect that the goal of parsimoniously accounting for all of these findings will be a significant challenge for years to come.

4.2.7. Towards a more descriptive theory of causal judgments

A crude summary of the human response patterns of judgments from reasoning on simple binary event causal networks would be that about half of the judgments are consistent with the simple probabilistic, Bayes Networks point-estimate normative calculations (e.g., .47 global r^2 between model and human responses). This simple fact provides a prima facie case for starting the development of a descriptive model from the basic Causal Bayesian Network model.

Out of all the models we presented based on the standard Bayesian Network model, we believe that it makes sense to start building a descriptive model by extending the Beta Inference Model. The peaks of the distributions predicted by the Beta Inference Model are equivalent to the standard point estimate Bayesian Network model, but one major advantage of the Beta Inference Model is that it also predicts variance and skew in judgments that were often borne out in the data (see Appendix B). Though the other versions of the Bayesian Network model (e.g., Shared Generative Cause) could also be extended to incorporate variance, there are multiple implementation details that would need to be specified (e.g., whether the network is parameterized as causal strengths or conditional probabilities), which further increases the complexity of searching for the one best combination of models in a principled way. In contrast, the Beta Inference Model is simple and already predicts variance and skew without any additions.

However, a fundamental problem is that the Beta Inference Model requires learning data to instantiate the α and β parameters of the model (this would also be a problem for other Bayesian models that predict variance based on learning data). But in most of the previous experiments learning data were not provided, and violations of the Markov Assumption and problems with explaining away were observed. One possible solution is to suppose that when participants do not have learning data, that they reason as though they were sampling experienced cases, but with the cases imagined or mentally simulated (analogous to the "mental models" proposed by Johnson-Laird and others, e.g., Goldvarg & Johnson-Laird, 2001; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999).

An even simpler account is that when participants do not have learning data, that their judgments are primarily based on simple heuristics like ambiguity aversion, conflict aversion, and a monotonicity assumption. It seems that any model that aspires to account for all of the systematic patterns in the current data must include a special explanation for the occasional spikes of responses at the center of the scale (.50), which is most easily explained by the simple heuristics. In the case of inferences on the three-event networks, it clearly appears that event-event conflicts (e.g., effect 1 occurs, effect 2 does not occur in a common-cause network) and missing information (the state of a mediating event is unknown in a causal-chain network) are likely to cue participants to respond with the ".50" value to express their high level of uncertainty. In situations without learning data, it is plausible that these heuristics would play an even larger role.

In sum, though we acknowledge that there are other possible combinations of models, the Beta Inference Model plus some simple heuristics capture most of the phenomena (means, variance, and skew) observed in our experiments.

4.2.8. Value of the current work

Despite the challenging theoretical situation we find ourselves in, we believe that the current findings and the comparison of theories in Table 11 are valuable for a number of reasons. First, pointing out a larger set of findings that must be accounted for is an advance in and of itself. This includes identifying other judgments that are outliers such as inferences on the middle variable on a common effect, as well as scrutinizing the distributions of judgments for previously-hidden spikes. Assembling these findings provides an empirical base from which to move forward.

Second, Table 11 assembles a large number of theories that can be considered in the future. Furthermore, we have revealed a number of ways that specific theories cannot simultaneously account for two patterns, which helps to reveal the limitations of the individual theories.

Third, we introduced a number of new theories including the Beta Inference Model and three simple heuristics. Given the utility of the heuristics and biases approach for judgment and decision making more broadly, we suspect that a heuristics approach for causal judgment may be fruitful.

We now return to a discussion of two especially important behavioral findings.

4.3. Explaining away

These experiments are the first to demonstrate insufficient explaining away after learning from case-by-case experiences with the multivariate distribution, and the finding is bolstered by consistent patterns for both binary and numerical variables. These findings are also consistent with a broader pattern of insufficient explaining away in other experiments that involved no experiential learning (Rehder, 2014; Sussman & Oppenheimer, 2011).

The reason for insufficient explaining away is not entirely clear. Associative or “spreading activation” reasoning is one possible mechanism (Rehder, 2014). The idea behind this explanation is that the structure $[X_1 \rightarrow Y \leftarrow X_2]$ is represented as a $[X_1 - Y - X_2]$, in which case X_1 and X_2 are unconditionally correlated. Another explanation is that perhaps people have difficulty learning the statistical relations between the three variables. Table 1 shows the probability of X_1 given Y and X_2. The interesting feature of this structure is that when inferring X_1, there is a strong interaction between Y and X_2. In contrast, when inferring Y given X_1 and X_2, both X_1 and X_2 have main effects, potentially making it easier to learn how to predict Y than to predict X. Future research can investigate whether the statistical properties of common effect structures contribute to the difficulty of explaining away.

Despite the insufficient explaining away observed here, we believe that in certain concrete everyday situations people do understand the explaining away principle. When people hear about familiar social and medical situations where explaining away should apply, they show a grasp of the concept, although we suspect they do not have a precise understanding of the quantitative relationships.

Three possible factors present in everyday reasoning might support explaining away inferences. First, explaining away seems to appear when reasoning about very rare and or very strong causes (McClure, 1998). Second, explaining away might be facilitated by reasoning about concrete mechanisms and inhibited by probabilistic reasoning (Ahn & Bailenson, 1996; Park & Sloman, 2013). Third, explaining away may arise due to domain-specific heuristics or causal rules (cf. pragmatic reasoning schemas Cheng & Holyoak, 1985). For example, a doctor might conceptualize two rare diseases as mutually exclusive. Fourth, it may be facilitated by reasoning about obvious causal relations from prior knowledge (Oppenheimer, Tenenbaum, & Krynski, 2013) rather than novel relations from experience. In sum, there is much work yet to understand when people may appropriately explain away.

4.4. Violations of the Markov Assumption

Despite the many studies on violations of the Markov Assumption, this is only the second time that the Markov Assumption has been tested after participants have learned the probabilistic relations between the variables from experience (cf. Park & Sloman, 2013, Experiment 3). When individuals are asked to reason about a causal structure but are not given experience, they may not fully grasp what the causal structure implies about the probabilistic relations, and consequently, it is less surprising that they fail to understand the Markov Assumption implicit in the structure.

However, when participants are given experience with data that exemplifies the Markov Assumption, the continued violation of the assumption, in our opinion, is more reason for concern. It is not that the participants did not learn from the data at all; when comparing Experiment 1 versus Experiment 2 participants were clearly sensitive to the parameters inherent in the data. Furthermore, Experiment 1b had 128 trials, which is a substantial sample of data. These findings raise the future question of what sort of training or experience must participants receive in order to not violate the Markov Assumption.

4.5. Induction versus deduction

Previous studies examining the 'explaining away' phenomenon and the violation of the Markov Assumption primarily concern deductive causal reasoning.[12] Participants were typically given verbal information on various causal structures and asked to answer questions, where the answers were deducible from the given information. In contrast, the experiments in this paper presented participants with trial-by-trial data, and thus may appear to involve inductive causal reasoning. To elaborate on this thought, the most common distinction between deduction versus induction is that deductive arguments are intended to be valid – the truth of the conclusion is guaranteed by the truth of assumptions. In contrast, inductive arguments are meant to be probabilistic – the conclusion is stronger based on the amount of the evidence (The Internet Encyclopedia of Philosophy, 2016).

There are at least two reasons why the current task might be viewed as inductive, whereas the tasks in previous studies might be viewed as deductive. The first is that participants were presented learning data. The second is that the judgments were quantitative, which presumably encouraged use of the experienced data and also encouraged participants to view the judgments not as valid but supported, to varying degrees, based on the data. Of course, the distinction between induction versus deduction is tricky and it is not always easy to classify a particular judgment or argument as one or the other. In a sense, what truly determines whether the judgment was made inductively or deductively is the mental process that was underlies it. For example, participants could infer the parameters from the learning data, and then deduce an inference (the typical presentation of the CBN theory), they could make the judgment straight from the learning data (e.g., the Beta Inference Model), or they could deduce a judgment based solely on a heuristic (e.g., ambiguity aversion, monotonicity assumption). Our best guess is that a variety of different inductive and deductive processes were used.

4.6. Limitations

The current research has a couple of important limitations. First, this research only studied predictions of one variable given one or two other variables. In contrast, one of the most important aspects of the CBN theory is that it explains how to make inferences after interventions (Meder et al., 2008; Rips, 2010; Sloman & Lagnado, 2005; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). However, all of the theories discussed in Table 11 can be applied for interventions. Intervention judgments can be viewed as predictions made on a modified ('severed') causal structure. The heuristic models could be amended so that the only variables that are used in the judgment are those that are still connected to the to-be-inferred variable after the structure is modified. The Beta inference Model could also be modified for making intervention judgments (see Appendix B). In sum, it is likely that many of the current findings will also be relevant for intervention judgments.

Second, though we have argued that providing learning data is important, an argument can be made against learning data. Specifically, it is unclear whether participants use learning data to augment their representation of the causal structure (in line with CBN theory), or whether they view the learning data and the causal structure as independent sources of information that they use to make inferences. If the latter is true, then the paradigm that includes a structure and learning data involves studying two processes simultaneously, neither of which is well understood, so it would

[12] We thank an anonymous reviewer for this insight.

be difficult to attribute any results to one representation, the other, or an interactive process. For one example, having learning data introduces the possibility of other complexities like order effects.[13]

We agree that introducing learning data does complicate the scientific processes. On the other hand, not having learning data also complicates the process in other respects. For example, if participants think that the causal strengths are weak, the normative amount of explaining away can be miniscule, so not finding any explaining away in those tasks could be interpreted as normative. Or for another example, subjects may simply not understand the formalism of how the arrows and nodes that comprise a network imply conditional independencies through the Markov condition, even if they could understand the Markov condition through better training or through learning data. We also believe that though some real-world judgments are made purely from causal structure knowledge, many involve some combination of knowledge of a structure as well as some experience (e.g., when deciding whether to study an additional hour for a test, a learner has prior experiences with amount of studying and test outcomes). In sum, we believe that both approaches are important for better understanding causal inference.

A third limitation is that our research only investigated a limited range of the parameter space. We could have tested cases with extremely strong or extremely weak causal relations, as well as rare causes. Instead we chose to study moderately to fairly strong causal relations, and base rates of .5 (.25 for the common effect in Experiment 1b). These choices were made because we wanted to choose fairly neutral parameters for this initial investigation, and because changing away from neutral parameters requires increasing numbers of learning trials to accurately instantiate the parameters. We can speculate on how the judgments might change with other parameters. We hypothesize that if causal strengths are weaker, that it will be harder to detect the weakness effects – the typically weak inferences will come into line with the normatively weak inferences. We suspect that the too-strong inferences (e.g., inferring the effect in the common effect structure) will remain too strong. In contrast, if the strengths are increased, we expect more inferences to be weaker than normative. We expect the amount of explaining away to increase (as it should normatively; cf. Experiment 1a versus 1b), though we still expect that the amount of explaining away to be too small relative to the normative amount. We do not have predictions for the Markov assumption.

4.7. Conclusion

In the last decade the interest in causal probabilistic graphical models has produced a surge of research on whether humans approximate the normative calculations implied by these models. The present research provides a comprehensive examination of causal reasoning in three-event scenarios, representing chain, common cause, and common effect structures, looking at almost all of the possible inferences, on both binary and numerical variables.

We conclude that even though the point-estimate normative model captures the overall trends of judgments at a high level, individual judgments deviate from the model in systematic ways. In particular, our studies show that violations of the Markov Assumption usually persist even when participants have access to learning data, inferences are usually too weak (with a few local exceptions), and explaining away judgments often fail to show the qualitative explaining away pattern. And, even when they do show the qualitative pattern, they are typically not sufficiently strong quantitatively. These results, as well as a qualitative analysis of 12 theories to explain the results, and help guide the next generation of research and theory on causal reasoning.

Appendix A. Judgments broken down more finely

In the main manuscript, symmetric inferences (e.g., inferences on X_1 and X_2) were collapsed, and inferences on the lower half of the scale were flipped to the upper half. Here we break out these judgments for the Markov Assumption and Transitive inferences in Experiment 1. There are no important and consistent trends to report.

[13] We thank Bob Rehder for this insight.

A.1. Markov Assumption inferences in Experiment 1

Inference type	Judgment	
	Exp. 1a	Exp. 1b
Chain: $X_1 \rightarrow Y \rightarrow X_2$		
$P(x_1 = 1 \mid y = 1, x_2 = 1)$	80	83
$P(x_1 = 1 \mid y = 1, x_2 = 0)$	59	71
$P(x_1 = 1 \mid y = 0, x_2 = 1)$	42	31
$P(x_1 = 1 \mid y = 0, x_2 = 0)$	21	21
$P(x_2 = 1 \mid y = 1, x_1 = 1)$	79	83
$P(x_2 = 1 \mid y = 1, x_1 = 0)$	62	66
$P(x_2 = 1 \mid y = 0, x_1 = 1)$	42	33
$P(x_2 = 1 \mid y = 0, x_1 = 0)$	25	21
Common cause: $X_1 \leftarrow Y \rightarrow X_2$		
$P(x_1 = 1 \mid y = 1, x_2 = 1)$	74	80
$P(x_1 = 1 \mid y = 1, x_2 = 0)$	66	73
$P(x_1 = 1 \mid y = 0, x_2 = 1)$	35	29
$P(x_1 = 1 \mid y = 0, x_2 = 0)$	19	20
$P(x_2 = 1 \mid y = 1, x_1 = 1)$	78	80
$P(x_2 = 1 \mid y = 1, x_1 = 0)$	65	71
$P(x_2 = 1 \mid y = 0, x_1 = 1)$	35	25
$P(x_2 = 1 \mid y = 0, x_1 = 0)$	25	20

A.2. Transitive inferences in Experiment 1

Inference type	Judgment	
	Exp. 1a	Exp. 1b
Chain: $X_1 \rightarrow Y \rightarrow X_2$		
$P(x_2 = 1 \mid x_1 = 1)$	59	69
$P(x_2 = 1 \mid x_1 = 0)$	40	32
$P(x_1 = 1 \mid x_2 = 1)$	58	71
$P(x_1 = 1 \mid x_2 = 0)$	37	32
Common cause: $X_1 \leftarrow Y \rightarrow X_2$		
$P(x_2 = 1 \mid x_1 = 1)$	60	66
$P(x_2 = 1 \mid x_1 = 0)$	35	33
$P(x_1 = 1 \mid x_2 = 1)$	60	66
$P(x_1 = 1 \mid x_2 = 0)$	40	31

Appendix B. The Beta Inference Model

Appendix B re-presents Figs. B.1–B.4 from the main manuscript, with predictions from the Beta Inference Model shown in dashed curves. The peaks of the Beta Inference Model curves always equal the normative point estimate (the vertical line), but the Beta Inference Model often captures some of the variance and asymmetries seen in the judgments.

At the end of the appendix we discuss how the Beta Inference Model could be used to make judgments from interventions instead of observations.

B.1. Violations of the Markov Assumption in Experiment 1

The Beta Inference Model captures the basic finding of violations of the Markov Assumption because the distribution for the "low" inferences is more skewed with more weight towards .50 than the distribution for the "high" inferences. The model also captures another finding; there is no violation of the Markov Assumption for the common effect for Experiment 1a, but there is for Experiment

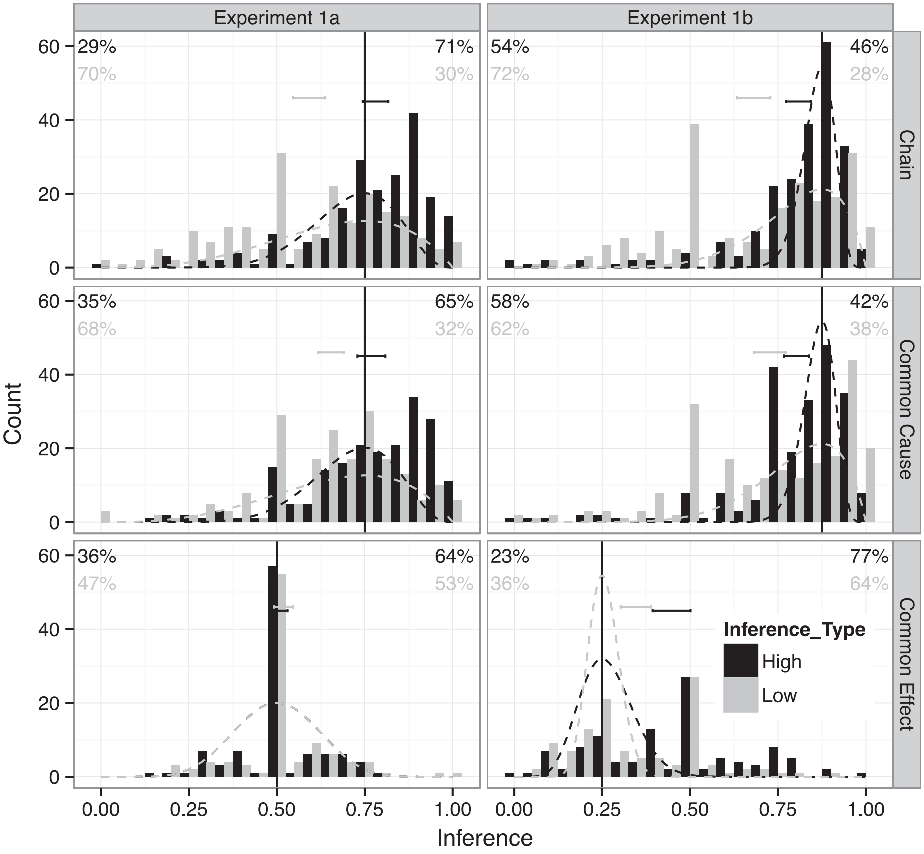


Fig. B.1. Distributions of individual responses showing Markov Violations from Experiment 1. Note: This is same as Fig. 3 in the main article with the addition of the dashed curves, which represent the predictions of the Beta Inference Model. Gray bars indicate responses when the screened-off variable has a low value, darker bars when the screened-off variable has a high value (the two sets of bars should be identical, if the Markov Assumption holds). The thin vertical line in each panel represents the 'correct' normative point-estimate response. The horizontal bars high up in each panel represent 95% confidence intervals on the means for each condition within the panel computed from regressions with by-subject random effects on the intercept. The numbers are the percent of inferences on either side of the normative calculation, after removing judgments that are exactly correct.

1b. The main weakness is that the model does not capture the spikes at .50, which would necessitate an additional rounding assumption for judgments close to .50.

B.2. Transitive inferences in Experiment 1

The Beta Inference Model captures roughly the variability around the single point estimate inference, and because of the skew in the distributions it also captures some of the weakness of the judgments. However, the spikes at .50 would require an additional rounding assumption.

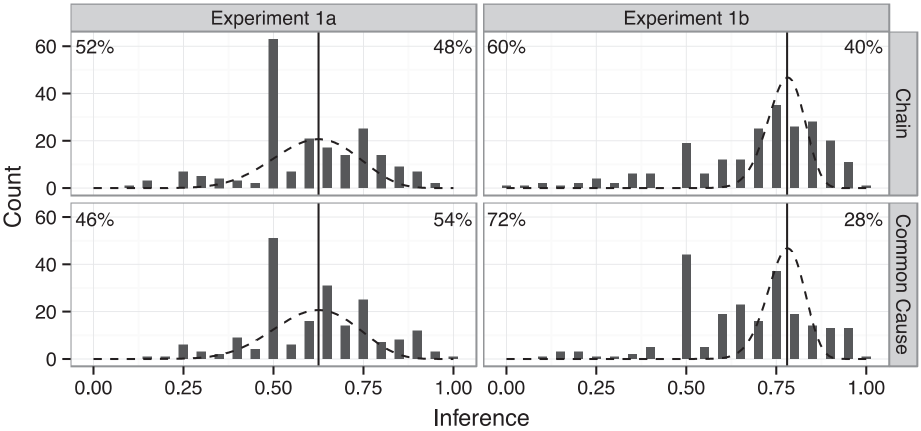


Fig. B.2. Distributions of responses to questions testing transitive inferences in Experiment 1. Note: This is same as Fig. 4 in the main article with the addition of the dashed curves, which represent the predictions of the Beta Inference Model. The thin vertical line in each panel represents the 'correct' point-estimate normative response. The numbers are the percent of inferences on either side of the normative calculation, after removing judgments that are exactly correct.

B.3. Inferences to "middle" variables in Experiment 1

The Beta Inference Model does a fairly good job of capturing the distributions of the judgments, and also captures the weakness of the inferences of the Chain, Common Cause, and Common Effect Low judgments through the skew of the distributions. The model does not capture the Common Effect High judgments.

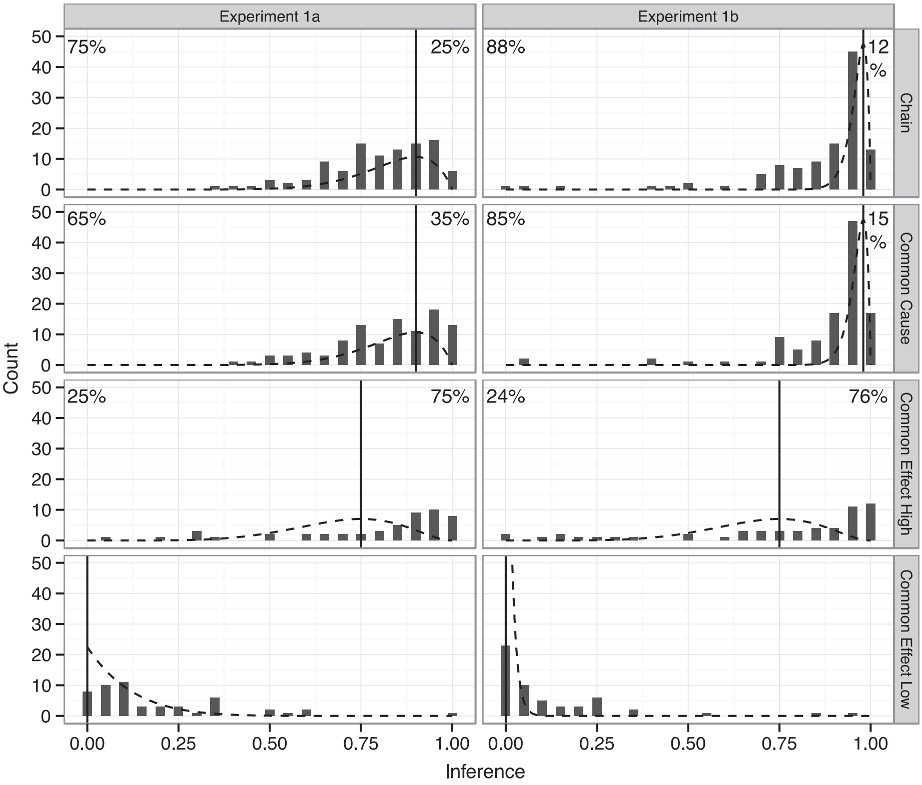


Fig. B.3. Inferences to "Middle" Variables in Experiment 1. Note: This is same as Fig. 5 in the main article with the addition of the dashed curves, which represent the predictions of the Beta Inference Model. The normative common effect low $P(y = 1|x_i = 0, x_j = 0)$ inferences are not equal to the normative common effect high $P(y = 1|x_i = 1, x_j = 1)$ inferences, so they are not flipped to the upper end of the scale and are presented separately. The thin vertical line in each panel represents the 'correct' normative response. The numbers are the percent of inferences on either side of the normative calculation, after removing judgments that are exactly correct.

B.4. Explaining away in Experiment 1

The skew in the Beta Inference Model captures the weak explaining away inferences to a certain extent – most prominently for the $P(x_i = 1|y = 1, x_j = 0)$ inference. However, it does not explain why the $P(x_i = 1|y = 1, x_j = 1)$ judgments are so high. The model captures some of the skew in the $P(x_i = 1|y = 1)$ judgments, but the spike at .50 would necessitate an additional assumption.

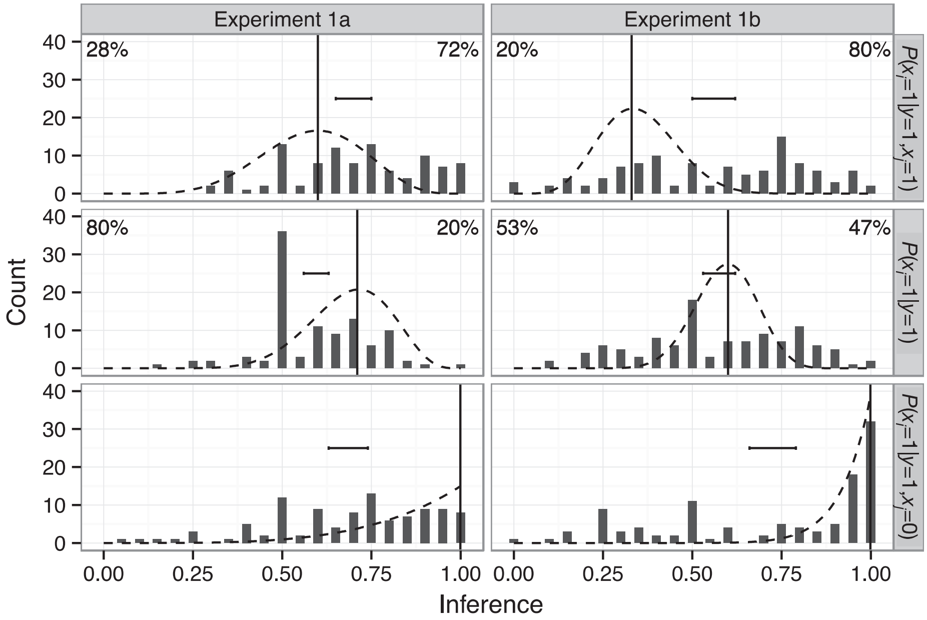


Fig. B.4. Distributions of individual responses that tested "explaining away" in Experiment 1 on common effect [$X_1 \rightarrow Y \leftarrow X_2$]. Note: This is same as Fig. 6 in the main article with the addition of the dashed curves, which represent the predictions of the Beta Inference Model. Vertical lines are the normative answers. Horizontal bars are 95% CI of the mean inference. The numbers are the percent of inferences on either side of the normative calculation, after removing judgments that are exactly correct.

B.5. Judgments from interventions

One of the most important aspects of the CBN theory of causal reasoning is that it differentiates between observations versus interventions (see Section 4.6 for citations). The core idea is that an intervention 'severs' the intervened-upon variable from that variable's causes, producing a new modified causal structure. In this section we discuss how the Beta Inference Model could handle interventions.

Certain interventions are trivial to accommodate with the Beta Inference Model. Inferences that only involve reasoning downwards in the direction of causality are the same for interventions and observations. For example, $P(x_2|do\ y = 1) = P(x_2|y = 1)$ on the chain [$X_1 \rightarrow Y \rightarrow X_2$], and $P(x_2|y = 1)$ can be easily calculated with the Beta Inference Model. Interventions that involve reasoning upwards against the direction of causality just involve base rates, which can also be inferred with the Beta Inference Model. For example, $P(x_1|do\ y = 1) = P(x_1)$ on the chain [$X_1 \rightarrow Y \rightarrow X_2$].

When the causal structure is more complex, interventions become more complicated to calculate, though the Beta Inference Model can still be applied. Consider the diamond causal structures studied by Meder, Hagmayer, and Waldmann (2009) in which there are two paths from A to D; $A \rightarrow B \rightarrow D$ and $A \rightarrow C \rightarrow D$. Meder et al. showed that an inference $P(d = 1|do\ c = 1)$ should be calculated with the following expression:

$$P(a=1)P(b=1|a=1)P(d=1|b=1,c=1) + P(a=1)P(b=0|a=1)P(d=1|b=0,c=1)$$
$$+ P(a=0)P(b=1|a=0)P(d=1|b=1,c=1) + P(a=0)P(b=0|a=1)P(d=1|b=0,c=1)$$

Based on the topology of this graph, when C is intervened upon, A can be ignored entirely, so the above equation can be simplified to the following expression:

$$P(b=1)P(d=1|b=1,c=1) + P(b=0)P(d=1|b=0,c=1)$$

Then, each of the four probabilities in the above expression can be estimated using the Beta Inference Model explained in the main text. In the expressions below, $N(k)$ refers to the number of trials for which k holds true, explained in the main text.

$$P(b=1) = \mathrm{Beta}(N(b=1)+1, N(b=0)+1)$$

$$P(b=0) = \mathrm{Beta}(N(b=0)+1, N(b=1)+1)$$

$$P(d=1|b=1,c=1) = \mathrm{Beta}(N(d=1,b=1,c=1)+1, N(d=0,b=1,c=1)+1)$$

$$P(d=1|b=0,c=1) = \mathrm{Beta}(N(d=1,b=0,c=1)+1, N(d=0,b=0,c=1)+1)$$

The table below shows means of the results of 10,000 simulations of the Beta Inference Model for Meder, Hagmayer, and Waldmann (2009) $A_{high}C_{low}$ condition in Experiment 2. This condition was chosen because it best discriminates the observation and intervention predictions. The main point is that the Beta Inference Model comes fairly close to CBN for these judgments and captures the difference between interventions versus observations.

Inference	CBN	Beta Inference Model	Mean of subjects' judgments
$P(d=1\|c=1)$	.92	.90	.80
$P(d=1\|c=0)$	.23	.25	.39
$P(d=1\|do\ c=1)$	.88	.84	.77
$P(d=1\|do\ c=0)$	.50	.47	.47

References

Ahn, W., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology, 31*, 82–123. http://dx.doi.org/10.1006/cogp.1996.0013.

Anderson, N. H. (1981). *Foundations of information integration theory*.New York: Academic Press.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *The Behavioral and Brain Sciences, 30*, 241–254. http://dx.doi.org/10.1017/S0140525X07001653. discussion 255–297.

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive consumer choice processes. *Journal of Consumer Research, 25*, 187–217.

Burnett, R. C. (2004). *Inference from complex causal models*. Northwestern University.

Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty, 5*, 325–370. http://dx.doi.org/10.1007/BF00122575.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*.New York: Guilford Press.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367–405. http://dx.doi.org/10.1037//0033-295X.104.2.367.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*, 391–416. http://dx.doi.org/10.1016/0010-0285(85)90014-3.

Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review, 112*(3), 694–707. http://dx.doi.org/10.1037/0033-295X.112.3.694.

Christensen-Szalanski, J. J., & Beach, L. R. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance, 29*, 270–278.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106.

Eddy, D. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, UK: Cambridge University Press.

Edgell, S. E., Harbison, J. I., Neace, W. P., Nahinsky, I. D., & Lajoie, A. S. (2004). What is learned from experience in a probabilistic environment? *Journal of Behavioral Decision Making, 17*, 213–229. http://dx.doi.org/10.1002/bdm.471.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science, 21*, 329–336. http://dx.doi.org/10.1177/0956797610361430.

Fernbach, P. M., & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument & Computation, 4*, 64–88. http://dx.doi.org/10.1080/19462166.2012.682655.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*.Cambridge, UK: Cambridge University Press.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451–482.

Goedert, K. M., Harsch, J., & Spellman, B. A. (2005). Discounting and conditionalization: Dissociable cognitive processes in human causal inference. *Psychological Science, 16*(8), 590–595. http://dx.doi.org/10.1111/j.1467-9280.2005.01580.x.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25*, 565–610. http://dx.doi.org/10.1207/s15516709cog2504_3.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology, 166*, 129–166.

Hadar, L., & Fox, C. (2009). Information asymmetry in decision from description versus decision from experience. *Judgment and Decision Making, 4*, 317–325.

Hagmayer, Y., & Osman, M. (2012). From colliding billiard balls to colluding desperate housewives: Causal Bayes nets as rational models of everyday causal reasoning. *Synthese, 189*, 17–28. http://dx.doi.org/10.1007/s11229-012-0162-3.

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of their choices as interventions. *Journal of Experimental Psychology: General, 138*, 22–38. http://dx.doi.org/10.1037/a0014585.

Hastie, R. (2016). Causal thinking in judgments. In *Blackwell handbook of judgment and decision making*, pp. 590–628). New York: Blackwell.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 301–354). Springer.

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences, 13*, 517–523. http://dx.doi.org/10.1016/j.tics.2009.09.004.

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review, 114*, 733–758. http://dx.doi.org/10.1037/0033-295X.114.3.733.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology, 62*, 135–163. http://dx.doi.org/10.1146/annurev.psych.121208.131634.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. P. (1999). Naïve probability: A mental model theory of extensional reasoning. *Psychological Review, 106*.

Jones, E. (1979). The rocky road from acts to dispositions. *The American Psychologist, 34*, 107–117.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 66*, 497–527.

Kelley, H. H. (1972). Causal schemata and the attribution process. In E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 151–174). Morristown, NJ: General Learning Press.

Khemlani, S. S., & Oppenheimer, D. M. (2011). When one model casts doubt on another: A levels-of-analysis approach to causal discounting. *Psychological Bulletin, 137*(2), 195–210. http://dx.doi.org/10.1037/a0021809.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*, 1. http://dx.doi.org/10.1017/S0140525X00041157.

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

Kruschke, J. K. (2011). *Doing Bayesian data analysis*.Oxford: Academic Press.

Krynski, T. R., & Tenenbaum, J. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 136*, 430–450. http://dx.doi.org/10.1037/0096-3445.136.3.430.

Lichtenstein, S., Earle, T., & Slovic, P. (1975). Cue utilization in a numerical prediction task. *Journal of Experimental Psychology: Human Perception and Performance, 104*, 77–85.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*, 955–984. http://dx.doi.org/10.1037/a0013256.

Mayrhofer, R., & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science, 39*, 65–95. http://dx.doi.org/10.1111/cogs.12132.

McClure, J. (1998). Discounting causes of behavior: Are two reasons better than one? *Journal of Personality and Social Psychology, 74*, 7–20.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review, 15*, 75–80. http://dx.doi.org/10.3758/PBR.15.1.75.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition, 37*, 249–264. http://dx.doi.org/10.3758/MC.37.3.249.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review, 121*, 277–301. http://dx.doi.org/10.1037/a0035944.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117*, 68–85.

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. http://dx.doi.org/10.1037/0033-295X.102.2.331.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*.Cambridge, MA: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.

Neapolitan, R. E. (2004). *Learning Bayesian networks*.Upper Saddle River, NJ: Pearson Prentice Hall.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*.Englewood Cliffs, NJ: Prentice-Hall.

Oppenheimer, D. M., Tenenbaum, J., & Krynski, T. R. (2013). Categorization as causal explanation. Discounting and augmenting in a bayesian framework. *Psychology of learning and motivation – Advances in research and theory* (Vol. 58, pp. 203–231). Elsevier. http://dx.doi.org/10.1016/B978-0-12-407237-4.00006-2.

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology, 67*, 186–216. http://dx.doi.org/10.1016/j.cogpsych.2013.09.002.

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

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72*, 346–354. Retrieved from<http://www.ncbi.nlm.nih.gov/pubmed/5968681>.

Rehder, B. (2010). Causal-based categorization: State of the art. *The psychology of learning and motivation* (Vol. 52, pp. 39–116). Elsevier. http://dx.doi.org/10.1016/S0079-7421(10)52002-4.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology, 72*. http://dx.doi.org/10.1016/j.cogpsych.2014.02.002.

Rehder, B. (2015). The role of functional form in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 670.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*, 264–314. http://dx.doi.org/10.1016/j.cogpsych.2004.09.002.

Reips, U.-D., & Waldmann, M. R. (2008). When learning order affects sensitivity to base rates. *Experimental Psychology, 55*, 9–22. http://dx.doi.org/10.1027/1618-3169.55.1.9.

Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science, 34*, 175–221. http://dx.doi.org/10.1111/j.1551-6709.2009.01080.x.

Rips, L. J. (2008). Causal thinking. In J. E. Adler & L. J. Rips (Eds.), *Reasoning: Studies of human inference and its foundation* (pp. 597–631). Cambridge, UK: Cambridge University Press.

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140*, 109–139. http://dx.doi.org/10.1037/a0031903.

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science, 29*, 5–39. http://dx.doi.org/10.1207/s15516709cog2901_2.

Sloman, S. A., & Lagnado, D. A. (2015). Causality in thought. *Annual Review of Psychology, 66*, 223–247.

Steyvers, M., Tenenbaum, J., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27*, 453–489. http://dx.doi.org/10.1016/S0364-0213(03)00010-7.

Sussman, A., & Oppenheimer, D. (2011). A causal model theory of judgment. In C. Hölscher, Carlson, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the Cognitive Science Society* (pp. 1703–1708). Austin, TX: Cognitive Science Society.

The Internet Encyclopedia of Philosophy (2016). Deductive and inductive arguments. In *The Internet Encyclopedia of Philosophy*. . Retrieved from<http://www.iep.utm.edu/>.

von Sydow, M., Meder, B., & Hagmayer, Y. (2009). A transitivity heuristic of probabilistic causal reasoning. In N. A. Taatgen & H. van Rijn (Eds.). *Proceedings of the 31st annual conference of the Cognitive Science Society* (Vol. 1, pp. 803–808). Amsterdam: The Cognitive Science Society.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 216–227. http://dx.doi.org/10.1037/0278-7393.31.2.216.

Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In The. Oxford (Ed.), *Handbook of cognitive psychology* (pp. 733–752). Oxford, UK: Oxford University Press.

Yates, J., & Jagacinski, C. (1979). Nonregressiveness of subjective forecasts. In J. F. Hair, A. C. Burns, T. Oliva, & M. Peters (Eds.), *Proceedings of the 11th annual meeting of the annual meeting of the american institute for decision sciences* (pp. 132–134).

Yeung, S., & Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology, 76*, 1–29. http://dx.doi.org/10.1016/j.cogpsych.2014.11.001.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience, 6*, 1–14. http://dx.doi.org/10.3389/fnins.2012.00001.