Mem Cogn (2017) 45:245–260
DOI 10.3758/s13421-016-0662-3

Failures of explaining away and screening off in described versus experienced causal learning scenarios

Bob Rehder[1] · **Michael R. Waldmann**[2]

Published online: 8 November 2016

Abstract Causal Bayes nets capture many aspects of causal thinking that set them apart from purely associative reasoning. However, some central properties of this normative theory routinely violated. In tasks requiring an understanding of *explaining away* and *screening off,* subjects often deviate from these principles and manifest the operation of an associative bias that we refer to as the *rich-get-richer* principle. This research focuses on these two failures comparing tasks in which causal scenarios are merely described (via verbal statements of the causal relations) versus experienced (via samples of data that manifest the intervariable correlations implied by the causal relations). Our key finding is that we obtained stronger deviations from normative predictions in the described conditions that highlight the instructed causal model compared to those that presented data. This counterintuitive finding indicate that a theory of causal reasoning and learning needs to integrate normative principles with biases people hold about causal relations.

Keywords Causal reasoning · Causal learning · Explaining away · Reasoning errors · Markov violations

In the past two decades causal Bayes nets have emerged as the dominant theoretical tool to model complex causal reasoning. They represent causal knowledge as a set of variables that encode causes and effects and a set of causal arrows representing causal influences directed from causes to effects. For example, Fig. 1a presents a *common effect model* with two causes generating a joint effect. This model might represent, for example, that bacteria and viruses are two independent causes of fever. Figure 1b, by contrast, depicts a *common cause model* with one cause generating two effects. This model might represent that a virus causes two different symptoms. Both networks in Fig. 1 embody default assumptions regarding the independence of the underlying mechanisms responsible for the causal relations. In the common cause network, the cause independently generates the two effects. And, in the common effect network it is often assumed that the two generative causes operate independently, implying a *noisy-or* integration function in which each cause leads to an increase in the probability of the effect. Although more complex networks can be constructed we focus on those in Fig. 1 because they present reasoners with the simplest reasoning situations that nonetheless yield theoretically important predictions.

Research has shown that causal Bayes nets capture central features of human causal reasoning (see Rehder in press-a, in press-b; Rottman, in press; Rottman & Hastie, 2014; Waldmann, in press, Waldmann & Hagmayer, 2013; for overviews). Consistent with causal Bayes nets, people draw different inferences with common cause and common effect models, reflecting their sensitivity to causal direction (because those graphs are equivalent if one ignores the arrow heads; Rehder & Hastie, 2001; Waldmann, 2000; Waldmann & Holyoak, 1992; Waldmann, Holyoak, & Fratianne, 1995). In a common effect model, people know that the effect is more likely if more causes are present (although they can also reason with more complex integration functions when given reason to, such as when causes operate conjunctively; Griffiths, in press; Lucas & Griffiths, 2010; Rehder, 2014b; also see Waldmann, 2007). In a common

✉ Bob Rehder
bob.rehder@nyu.edu

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

[2] Department of Psychology, University of Göttingen, Göttingen, Germany

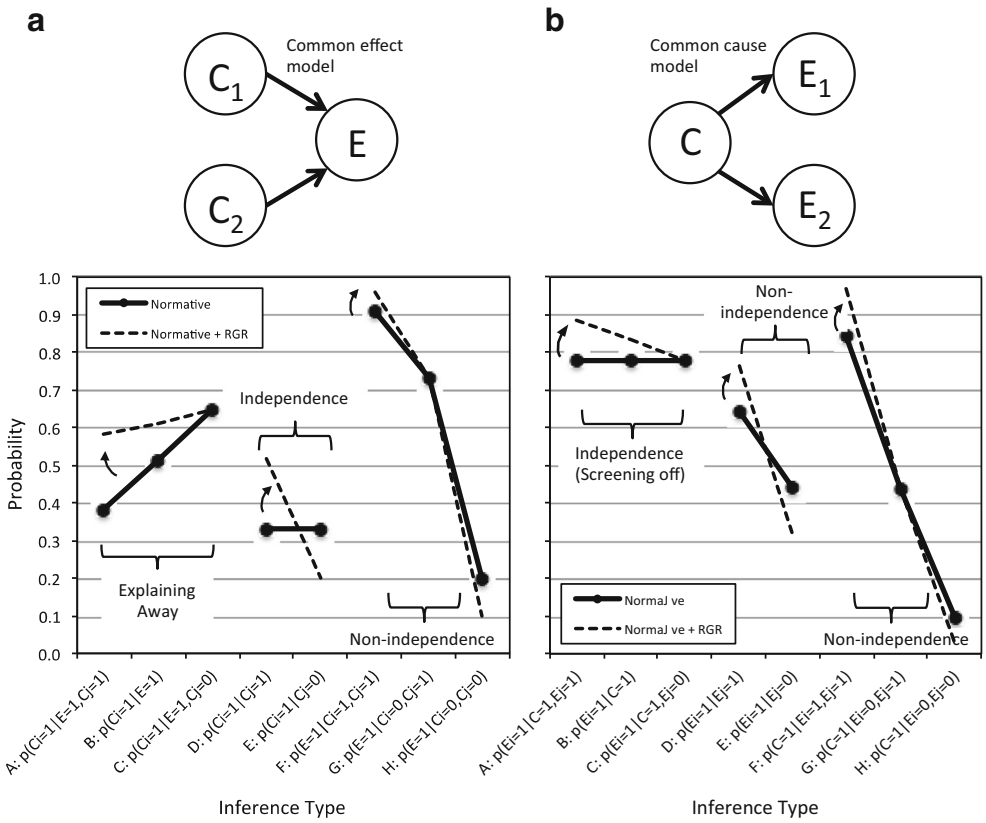

Fig. 1 (**a**) Normative predictions for a common effect network. (**b**) Normative predictions for a common cause network. The dashed lines represent the prediction of the normative model augmented with a rich-get-richer bias. Arrows represent the negative shift in line slopes that occurs with introduction of the rich-get-richer bias

cause model, they not only know that the cause is more likely if more effects are present but also that one effect implies another when the status of the cause is unknown (Rehder & Burnett, 2005). They know that *intervening* on a cause can potentially generate its effect but not vice versa (Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005). Causal reasoning can even reduce or eliminate standard reasoning fallacies (e.g., base rate neglect, Ajzen, 1977; Bar-Hillel, 1980; Hayes et al., 2014; Krynski & Tenenbaum, 2007; Tversky & Kahneman, 1980). As a result of these successes, causal models now play a key role in theories of conceptual structure (Kemp & Tenenbaum, 2009; Oppenheimer, Tenenbaum, & Krynski, 2013; Rehder, 2003a, 2003b, 2014; Rehder & Kim, 2009, 2010), inductive reasoning (Holyoak, Lee, & Lu 2010; Kemp, Shafto, & Tenenbaum, 2012; Lassaline, 1996; Lee & Holyoak, 2008; Rehder, 2006, 2009; Shafto, Kemp, Bonawitz, Coley, & Tenenbaum, 2008), decision making (Hagmayer & Sloman, 2009; Hagmayer & Meder, 2013), explanations (Lombrozo, 2010), and counterfactual reasoning (Pearl, 2000; Rips, 2010; Rips & Edwards, 2013).

Nevertheless, some of people's causal inferences are inconsistent with causal Bayes nets. For example, a common effect model entails the principle of *explaining away*, cases in which the presence of one cause in a common effect network makes another less likely. In fact, people explain away too little relative to the prediction of causal Bayes nets. Causal Bayes nets also embody *independence* constraints, cases in which variables should be probabilistically independent of one another. In fact, people's inferences often violate those independence constraints.

This article is organized as follows. The following two sections describe the principles of explaining away and independence and present the empirical evidence that people violate those principles. In the third section, we argue that those violations are manifestations of a common principle, one we refer to as the *rich-get-richer* principle. We then introduce a new experimental paradigm for assessing accounts of human causal reasoning, namely, one that builds on recent research on the *description-experience gap*. As in that research, we examine how inferences are drawn on the basis of causal relations that are (verbally) *described* as compared to those that are also (or instead) *experienced*, that is, manifested as a series of observations of individual cases that reflect the correlations implied by those causal relations. Besides being theoretically and practically important in its own right, we show that the described versus experienced comparison allows the predictions of the rich-get-richer principle to be assessed for new types of inference problems. These new predictions are then tested in two experiments.

Failures of explaining away

Explaining away is a signature property of common effect models with independent causes (Fig. 1a). Continuing our example, suppose a virus (C_1) and a bacterium (C_2) independently cause fever (E). If E is observed to occur, then the probability that, say, the virus C_1 is present increases. But if it is then further observed that the second cause C_2, the bacterium, is present, then the probability that virus C_1 is also present *decreases*. When an effect is present, the reduction of the probability of one cause when another is observed is called explaining away. The extent to which explaining away is normative depends on the strengths and necessity of the causes and the degree to which those causes are correlated (see Morris & Larrick, 1995; Rottman & Hastie, 2014).

In light of the central role of explaining away in causal reasoning, it is surprising that few studies have rigorously tested whether human reasoning honors this principle. One reason for this oversight may be that in social psychology a related phenomenon, *discounting* (see Jones, 1979; Kelley, 1972) has been observed in many studies so that explaining away has often been taken for granted (see Khemlani & Oppenheimer, 2010; McClure, 1998, for reviews). However, discounting has been used as an umbrella term for different phenomena. First, people may discount because they believe that the causes are mutually exclusive (or at least negatively correlated). For example, upon observing that a street is wet and a sprinkler is on, people tend to discount rain as a cause of the wetness (Pearl, 2000). But this inference may be due to prior knowledge that sprinklers are typically turned off when it rains (i.e., a negative correlation between sprinklers and raining) rather than explaining away. Second, people's *diagnostic* inferences (an inference from an effect to a cause, i.e., $p(C=1|E=1)$, where $C=1$ and $E=1$ denote that the cause and effect are present, respectively) are weaker when they are aware that E has alternative causes that are numerous and/or strong (e.g., Meder, Mayrhofer, & Waldmann 2014; Oppenheimer et al., 2013; Rehder & Kim, 2009; Waldmann, 2000; Waldmann & Hagmayer, 2005). For example, the probability of a disease given one of its symptoms is lower to the extent that the symptom can be caused by many other diseases. In this case discounting characterizes how people reason as a causal model is augmented with additional causal relations. A third case that is occasionally confused with discounting or explaining away is cue competition in learning, as predicted by the Rescorla–Wagner rule (Rescorla & Wagner, 1972), for example. Cue competition is different from explaining away as it is only operative during learning of associative weight parameters; its effect is a result of correlations between cues. By contrast, explaining away can also occur after learning when causes are independent and when both are strongly associated with the effect.

We adopt a stricter definition of explaining away that sets it apart from these other phenomena (e.g., negatively correlated causes and the existence of alternative causes). In particular, explaining away entails the inequality

$$p(C_i = 1|E = 1, C_j = 1) < p(C_i = 1|E = 1) < p(C_i = 1|E = 1, C_j = 0) \quad (1)$$

For example, if a bacterium and a virus independently cause a symptom, the probability of the bacterium given the symptom is lower if the virus is also present and higher if the virus is absent. The explaining away relationship is represented by the positively sloped solid line linking these inference types (labeled A, B, and C) in Fig. 1a.[1]

In fact, studies that have applied this stricter test of explaining away have yielded mixed results. Most experiments have either found that subjects explain away too little (Fernbach & Rehder, 2013; Morris & Larrick, 1995) or not at all; in some, the opposite result—an augmentation effect in which $p(C_i=1|E=1,C_j)>p(C_i=1|E=1,C_j=0)$ —was obtained (Fernbach & Rehder, 2013; Rehder, 2014a; see Rottman & Hastie, 2014, for a review).

Failures of independence

Another common failure is that reasoners are not always sensitive to the independence relations stipulated by causal Bayes nets. Specifically, when the state of a variable's direct causal parents is known, the *causal Markov condition* stipulates that that variable is conditionally independent of each of its nondescendants (Hausman & Woodward, 1999). This condition has a straightforward causal interpretation: Apart from its descendants, one has learned as much as possible about a variable once one knows the state of all of its direct causes.

For example, for the common cause network in Fig. 1b, the Markov condition stipulates that the two effects are independent conditioned on the cause, a condition commonly referred to as *screening off* (knowledge of C "screens off" the flow of information from one effect to another). That is, the following invariance (represented by

[1] The quantitative predictions in Figure 1A were generated assuming that the marginal probability of both C_1 and C_2 (i.e., their "base rates") is .33, the strength (or "causal power") of the both $C_1 \rightarrow E$ and $C_2 \rightarrow E$ is .67, and that the aggregate strength of alternative causes of E (i.e., causes other than C_1 and C_2) is .20. The predictions for the common cause network in Figure 1B are based on a marginal probability of C of .50, causal powers of .67, and alternative cause strength of .33. Note however the purpose of Figure 1 is to depict the qualitative pattern of inferences supported by the two types of networks, patterns that hold for all nondegenerate parameter values.

the solid horizontal line between problem types A, B, and C in Fig. 1b) should hold:

$$p(E_i = 1 | C = 1, E_j = 1) = p(E_i = 1 | C = 1) = p(E_i = 1 | C = 1, E_j = 0) \quad (2)$$

Numerous studies have shown that subjects who are asked to make an inference from a given cause C to one of its effects E_i, tend to be influenced by the presence or absence of other effects of C (Ali, Chater, & Oaksford, 2011; Mayrhofer & Waldmann, 2015; Park & Sloman, 2013; Rehder, 2014a; Rehder & Burnett, 2005; Rottman & Hastie, 2016; Walsh & Sloman, 2004; see Rottman & Hastie, 2014, for a review). For example, if a virus that is known to cause two symptoms is present in a particular patient, people tend to think that one symptom is more probable when the other symptom is present and less probable when it is absent.

The Markov condition also implies the existence of an independence relation in common effect models, namely, the causes should be independent when knowledge about the common effect is absent (represented by the solid horizontal line linking inference types D and E in Fig. 1a). This independence relation is also commonly violated: People often judge instead that $p(C_i = 1 | C_j = 1) > p(C_i = 1 | C_j = 0)$ (Perales, Catena, & Maldonado, 2004; Rehder, 2014a, 2014b; Rehder & Burnett, 2005; Rottman & Hastie, 2016).

Note that several theories accounting for such violations have been proposed in the literature (e.g., Park & Sloman, 2013, 2014). Some theories assume that people deviate from the instructed causal models by bringing to bear prior domain knowledge that leads them to augment the causal model. Others attribute the violations to the presence of more abstract domain knowledge (Mayrhofer & Waldmann, 2015; Rehder & Burnett, 2005; Waldmann & Mayrhofer, 2016). The General Discussion section will describe these proposals and evaluate them as potential accounts of our results.

The rich-get-richer principle

What is responsible for the weak (or nonexistent) explaining away and the violations of independence in human causal reasoning? Rehder (2014a) proposed that human causal reasoning exhibits an *associative bias* that subsumes both sorts of violations. This bias is synonymous with what we will call a *rich-get-richer* principle that states, in the case of causal models with generative links, that reasoners assume that one variable is more likely to be present to the extent that other variables in the causal model are also present. Conversely, the bias also entails that a variable is *less* likely to be present to the extent that other variables in the causal model are *absent* (the *poor-get-poorer* corollary of the rich-get-richer principle).[2]

It is important to distinguish our use of "association" from its use in traditional associative learning theory. In such theories, mechanisms are postulated that control how the strengths of acquired associations vary as a function of, for example, patterns of redundancy and competition between predictive cues (e.g., Rescorla & Wagner, 1972). By contrast, association in the present context refers to bidirectional noncompetitive relations that reflect the associations gleaned from the learning context.

The qualitative predictions of the rich-get-richer principle are shown in Fig. 1 as dashed lines superimposed on the normative predictions. First, consider explaining away in Fig. 1a. Although the normative model predicts that $p(C_i = 1 | E = 1, C_j = 1) < p(C_i = 1 | E = 1)$, the fact that two variables are present in the former scenario ($E = 1, C_j = 1$) as compared to only one in the latter ($E = 1$) raises the relative probability of C_i in the former scenario. Likewise, although the normative model predicts that $p(C_i = 1 | E = 1) < p(C_i = 1 | E = 1, C_j = 0)$, the fact that zero variables are absent in the former scenario ($E = 1$) as compared to one in the latter ($E = 1, C_j = 0$) raises the relative probability of C_i in the former. These qualitative predictions are shown as the dashed line connecting problem types A, B, and C in Fig. 1a. In particular, the slope of the line connecting these plot points has shifted to the negative (i.e., has become less positive).

The rich-get-richer principle also explains violations of independence. Recall that the causes of a common effect model should be unconditionally independent, $p(C_i = 1 | C_j = 1) = p(C_i = 1 | C_j = 0)$. But the rich-get-richer principle explains why C_i is viewed by reasoners as more probable in the former scenario (in which one variable, C_j, is present) as compared to the latter (C_j is absent). And, recall that the effects in a common cause model should be independent conditioned on the cause, $p(E_i = 1 | C = 1, E_j = 1) = p(E_i = 1 | C = 1) = p(E_i = 1 | C = 1, E_j = 0)$. But the rich-get-richer principle explains why E_i is viewed as most probable in the first scenario (two variables present) and least probable in the last scenario (one variable present, one absent). These qualitative predictions are shown as dashed lines for problem types D and E in Fig. 1a and types A, B, and C, in Fig. 1b, respectively. Again, the prediction is that the rich-get-richer effect will shift the slopes of the lines connecting these plot points to the negative.

As mentioned, one purpose of this article is to test the predictions of the rich-get-richer principle on new sets of problem types. One of these sets is types F, G, and H in Fig. 1a. The normative model predicts, unsurprisingly, that

[2] For the causal networks considered here, the rich-get-richer principle is equivalent to what Rottman and Hastie (2016) referred to as a *monotoncity principle,* in which the strength of causal inferences are a function of the number of variables present minus the number of variables absent.

the effect is more probable when more of its causes are present, that is, $p(E=1|C_i=1,\ C_j=1) > p(E=1|C_i=1,\ C_j=0) > p(E=1|C_i=0,\ C_j=0)$. A second set is types D and E in Fig. 1b, for which the normative model predicts that the two effects are unconditionally dependent, that is, $p(E_i=1|E_j=1) > p(E_i=1|E_j=0)$. A third set is types F, G, and H in Fig. 1b, for which the normative model predicts that the cause is more probable when more of its effects are present, that is, $p(C=1|E_i=1,\ E_j=1) > p(C=1|E_i=1,\ E_j=0) > p(C=1|E_i=0,\ E_j=0)$. The rich-get-richer principle predicts that all three of these effects should be stronger than predicted by the normative model, as shown by the negatively shifted dashed lines in Fig. 1. A contribution of the experiments that follow will be the establishment of a baseline condition—described in the following section—that establishes the existence of not only weak explaining away on independence violations but a negative shift in slope for all six inference sets in Fig. 1.

Note that our proposal is not that the rich-get-richer principle is the only influence on people's causal inferences. Rather, we claim that it functions as a bias that distorts the inferences implied by the normative model. Accordingly, the magnitude of the negative shifts in Fig. 1 will depend on how strongly the principle manifests itself in particular reasoning situations and particular reasoners.

Description versus experience in causal reasoning

Recent studies of judgment and decision making have uncovered fundamental differences between reasoning based on described versus experienced scenarios (see Hertwig, 2015, for an overview). For example, experiments on risky choice have asked if decisions depend on how the risk structure of gambles is conveyed—as a summary of payoff distributions or by having subjects experience the outcomes of a sequence of gambles. These studies revealed a *description-experience gap* in which sensitivity to outcomes with low probability depends on the presentation format (e.g., Hertwig, Barron, Weber, & Erev, 2004). Whereas initial research on the description-experience gap focused on gambles, more recently other domains have been studied (e.g., decision making in medical domains; Lejarraga, Pachur, Frey, & Hertwig, 2016).

Whereas causal Bayes nets have been tested with experientially conveyed statistical information (e.g., Gopnik et al., 2004; Griffiths & Tenenbaum, 2005; Meder et al., 2014; Rottman & Hastie, 2016; Waldmann, Holyoak, & Fratianne, 1995), many other studies have used verbal descriptions of scenarios to convey causal models and the strengths of the parameters (e.g., Fernbach, Darlow, & Sloman, 2011; Rehder & Hastie, 2001; Rehder & Kim, 2009). Few studies have been conducted that systematically compared different formats of presenting causal information. A typical finding of these reported by these studies is that causal strength is estimated differently depending on whether learning data are presented in trial-by-trial format or in a more compact format, for example, tabular summaries (see Perales, Catena, Cándido, & Maldonado, in press). In the description-based conditions, often no information is provided about quantitative parameters.

Our focus is on the influence of learning formats on subjects' ability to honor Markov constraints and to understand explaining away, which have only in one study been investigated by presenting causal models with trial-by-trial learning data (Rottman & Hastie, 2016). The majority of studies has presented causal models without data (e.g., (Ali et al., 2011; Mayrhofer & Waldmann, 2015; Park & Sloman, 2013, 2014; Rehder, 2014a; Rehder & Burnett, 2005; Walsh & Sloman, 2004). No systematic comparison between experience-based and description-based learning have been conducted targeting failures of causal reasoning in these very different formats of presenting causal information. One reason for this omission may be that the two paradigms were influenced by different research traditions. Whereas experience-based causal learning has been modeled after paradigms rooted in the associative learning literature (e.g., Cheng, 1997; Shanks & Dickinson, 1987; Waldmann & Holyoak, 1992), description-based learning paradigms are grounded in the tradition of theory-based categorization and conditional reasoning (Ali et al., 2011; Fernbach et al., 2011; Rehder & Hastie, 2001).

This study compares the two formats in tasks that focus on explaining away and independence. In the description-experience condition, subjects receive instructions about a causal model followed by a sample of data from the domain that informs them about the statistical relations between the causal events. Objective conditional probabilities computed on the basis of the data sample alone will exhibit, for example, explaining away and that the causes are unconditionally independent (in the common effect condition) and that the effects are independent conditioned on the cause (in the common cause condition). By contrast, the description-only condition provides only a description of the causal model. An experience-only condition presents the data, but not the causal model. This condition establishes a baseline regarding how reasoners draw inferences on the basis of the data sample.

A choice had to be made regarding how the data—the experience component—should be presented. In the description-experience gap literature, several formats have been tested ranging from sequences of gambles to lists of medical records, for example. In the causal learning literature, trial-by-trial learning has often been used, especially when causal theories were tested against alternatives from the associative learning literature (e.g., Waldmann, 2000). However, formats that avoid the influence of memory effects and reduce performance effects have also been tested (e.g., tabular information or compact graphic representations of cases; see Liljeholm & Cheng, 2007; Meder et al., 2014; Waldmann &

Hagmayer, 2001). These experiments use a format that corresponds to *record-based learning* in the judgment and decision-making literature in which statistical information is conveyed in a manner that is less cognitively demanding than in trial-by-trial learning.

What differences do we expect? The most extreme differences we expect are between the description-only and the experience-only conditions. In the description-only condition no data are presented so that inferences are based solely on causal model intuitions of the subjects. Thus, subjects are free to make assumptions about the strength of causal relations. In these kinds of situations, we typically see that subjects assume probabilistic, but fairly strong relations (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Mayrhofer & Waldmann, in press). Regardless of the assumed size of causal strength a failure of explaining away can be diagnosed from the inference patterns that they produce (see above). Because of the assumed rich-get-richer principle, we expect an attenuation of explaining away in this condition (see Rehder, 2014a), thus violating the normative implications of a causal Bayes net representation.

On the other end of the spectrum lies the experience-only condition. In this condition subjects are presented with tabulated data about the variables. No further cues (e.g., temporal order) that might suggest a specific underlying causal model are provided. Although it is possible to induce the class of Markov equivalent models from covariation data alone (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993), there is little evidence that untutored subjects are capable of doing this without the aid of computers. Most attempts to test their competency to induce causal models from covariation information alone have demonstrated poor performance despite the fact the tasks typically were simplified by restricting and prespecifying the causal models under consideration (e.g., Lagnado & Sloman, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Better performance can only be achieved with additional cues, such as temporal order or interventions (Lagnado, Waldmann, Hagmayer, & Sloman 2007). None of these cues were offered in the present experiments so that we expected that subjects in the experience-only condition would just read off the requested conditional probabilities from the data. Given that the data are normative with respect to the causal model, we therefore predict the strongest evidence for explaining away in this condition.

The most interesting condition is the description-experience condition. From the viewpoint of normative causal Bayes net theory, a theory that provides both normative data and a causal model whose structure is perfectly consistent with the data should lead to the best performance. Performance should be better than in the description-only condition because the data should help with getting the probability estimates right. And, given that explaining away is a hallmark of causal Bayes nets, graphically displayed causal models should further help. In light of the theoretical framework of causal Bayes nets, our prediction is therefore counterintuitive, especially when the contrast between experience-only and description-experience is considered. Although in both conditions objective data reflecting explaining away relations are readily available, we expect performance to be *worse* when additional information is given about the underlying causal model, as in the description-experience condition, relative to the experience-only condition. It will do so because causal models induce an associative bias (i.e., the rich-get-richer principle), one that will counteract the objective statistics in the data sample and shift the slope of the ABC line in Fig. 1a to the negative. We predict that the rich-get-richer principle will result in explaining away being the weakest of all in the description-only condition, in which the causal model is untethered from any objective data.

Analogous predictions apply to screening off. For a common cause model when the state of the cause is known, we predict that reasoners' tendency to incorrectly view the two effects as dependent will be smallest (indeed, nonexistent) in the experience-only condition and largest in the description-only condition. For a common effect model when the effect is unknown, the two causes will incorrectly be seen as dependent in the description-only condition, but not when the inferences are made on the basis of data alone. (Again, we expect performance in the description-experience condition to be intermediate between the description-only and experience-only condition.) That is, for the judgments that are the focus here, providing causal knowledge will make subjects' causal inferences *worse*. We even expect that the rich-get-richer bias will be observed in those inference types in Fig. 1 that are *not* independent (D and E in Fig. 1b and F–H in both figures). That is, we expect the slope of each line in Fig. 1 to exhibit a negative shift as causal model information is introduced.

Experiments 1 and 2

We tested subjects' inferences about a common effect model (Experiment 1) or a common cause model (Experiment 2). Each experiment contrasted the description-experience, description-only, and experience-only learning conditions, which we expected would moderate the predicted failures of causal model reasoning.

Method

Materials

Three domains were tested: economics, meteorology, and sociology. Subjects were first told that the domain they were about to study included three binary variables. For example, in the domain of economics they were told that *interest rates* could be either *low* or *normal*, *trade deficits* that were *small* or

normal, and *retirement savings* that were *high* or *normal*. In the domain of meteorology, the variables were ozone level, air pressure, and humidity; in sociology they were degree of urbanization, interest in religion, and socioeconomic mobility. To control for any domain knowledge that subjects might bring to the experiment, a four-level counterbalancing factor varied which senses of the variables were used. For example, depending on this counterbalancing factor, subjects in the economics condition learned that the nonnormal values for interest rates, trade deficits, and retirement savings were (low, small, high), (low, large, low), (high, small, low), or (high, large, high).

Experiments 1 and 2 tested common effect and common cause inferences, respectively. In both experiments, subjects in the description-only and description-experience conditions read two paragraphs describing two causal relations. The initial sentence of each paragraph stated that one variable caused another whereas the rest described the mechanism responsible for the causal relationship. For example: "Low interest rates cause high retirement savings. Low interest rates stimulate economic growth, leading to greater prosperity overall, and allowing more money to be saved for retirement in particular." The variable counterbalancing described above necessitated using different causal relationships, so that whereas some subjects were told that low interest rates cause high retirement savings, others were told that *high* interest rates cause high retirement savings, still others that low interest rates cause *low* retirement savings, and so forth. (See the Appendix for details about the variables, causal relationships, and their counterbalancing.) Subjects in both experiments were given additional instructions emphasizing the independence of the causal mechanisms. In Experiment 1 they were told "Remember that both C_1 and C_2 can each bring about E on its own. That is, it's not the case that both of these two have to be present for E to be present. Rather, C_1 can independently produce E on its own, and C_2 can independently produce E on its own as well" (where the experimenter used the variable names rather than C_1, C_2, and E). In Experiment 2 subjects were told "E_1 is a direct result of C, and E_2 is independently a direct result of C."

Subjects in the experience-only and description-experience conditions were given a data sheet depicting a sample of items drawn from the domain. In Experiment 1, the sample was the most likely one of size 27 drawn from a common effect model in which the base rate of the causes C_1 and C_2 is .32, the power of the causal links is .83, and the strength of alternative causes of E (the probability that E is present when C_1 and C_2 are both absent) is .12. The low base rates of C_1 and C_2 were chosen in order to yield a large normative explaining away effect (i.e., a substantial positive slope for the line linking inference types A, B, and C in Fig. 1a), thereby increasing the chance of observing a reduction in that slope as a consequence of the described causal model. In Experiment 2, the sample was the most likely one of size 33 drawn from a common cause model in which the base rate of the cause C is .50, the power of the causal links is .67, and the strength of alternative causes of E_1 and E_2 (the probability that either is present when C is absent) is .20. These resulting samples are presented in Table 1. Note that the sample sizes of 27 and 33 were chosen because they yielded samples whose statistics closely matched the target parameters.

The data sheet organized the sample items into subgroups that shared the same values for the variables. For example, Fig. 2a shows a subgroup of five economies that each have normal interest rates, small trade deficits, and high retirement savings. Each item in the subgroup was numbered (#6, #7, #13, etc.) to emphasize that it represented an individual economy. The data sheet showed eight such subgroups. The use of three domains and the four-level factor that varied which variable senses were used entailed the creation of 12 data sheets. In addition, we introduced a two-level factor in which two versions of each data sheet were created such that the subgroups appeared in different locations on the sheet. Thus, there were a total of 24 data sheets in each experiment.

Procedure

Participants first studied several computer screens of information about the domain and then performed the inference test. The initial screens presented a cover story and a description of the domain's three variables and their two values. Description-only and description-experience participants also observed screens that presented the two causal relationships and a diagram depicting the topology of the causal links (like those in Fig. 1). When ready, participants took a multiple-choice test of their knowledge. While taking the test, participants could return to the information screens they had studied; however, doing so obligated them to retake the test.

Table 1 The data samples presented to subjects in the description-experience and experience-only conditions of Experiments 1 and 2

Experiment 1				Experiment 2			
C_1	C_2	E	N	C	E_1	E_2	N
0	0	0	11	0	0	0	10
0	0	1	1	0	0	1	3
0	1	0	1	0	1	0	3
0	1	1	5	0	1	1	1
1	0	0	1	1	0	0	1
1	0	1	5	1	0	1	3
1	1	0	0	1	1	0	3
1	1	1	3	1	1	1	9

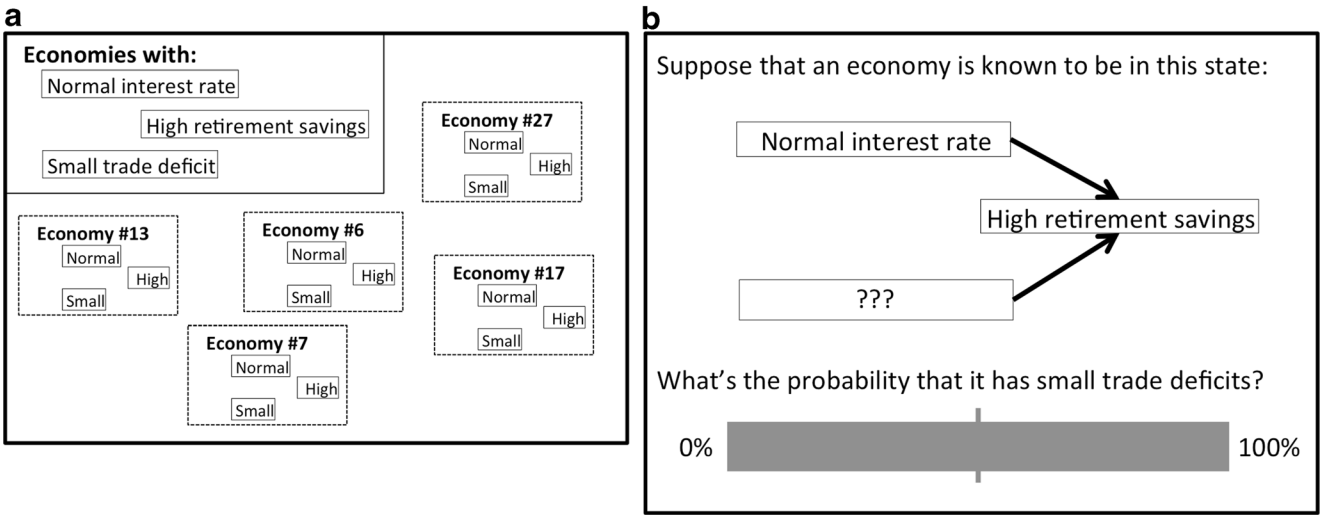

Fig. 2 (**a**) Example of a portion of one of the data sheets given to subjects in the economic condition. (**b**) Example of an inference question.

Subjects were then presented with the inference test. At the start of the test, subjects in the experience-only and description-experience conditions were provided a data sheet and told that "We have randomly chosen N [economies/societies/weather systems] from around the world and observed the three variables each has," with N = 27 (Experiment 1) or 33 (Experiment 2). Subjects in the experience-only condition were instructed to use the data sheet in answering the inference questions, those in the description-only condition were told to use the causal relations, and those in the description-experience condition were told to use both.

The test included the eight inference types (A–H) shown in Fig. 1a (Experiment 1) or 1b (Experiment 2). There were two versions of some inference types (A–E and G) in which C_1 and C_2 (Experiment 1) or E_1 and E_2 (Experiment 2) swapped roles. For instance, the two versions of inference type A in Experiment 1 were $p(C_2 = 1 | E = 1, C_1 = 1)$ and $p(C_1 = 1 | E = 1, C_2 = 1)$. For each inference, the known variable states were presented on the computer screen in boxes. Boxes associated with unknown variables were empty, with the exception of the to-be-predicted variable, in which case the box contained "???". For example, Fig. 2b depicts an economy known to have normal interest rates and high retirement savings and where trade deficits is the to-be-predicted variable. The boxes were connected with arrows that reflected the instructed causal relations in the description-only and description-experience conditions. There were no arrows in the experience-only condition in which no causal relations were instructed. Responses were entered by positioning a slider on a scale whose ends were labeled 0 and 100 % (see Fig. 2b). The order of the test questions was randomized for each participant. Inferences besides those in Fig. 1 were requested, but will not be discussed here.

Participants

Experiments 1 and 2 each tested 144 New York University undergraduates who received course credit for participating. In each, condition (experience-only, description-experience, description-only) was manipulated between subjects. In addition, each study used the two between-subjects counterbalancing factors described earlier (the four sets of variables senses and the two versions of the data sheets). Subjects were randomly assigned to these 3 × 4 × 2 = 24 between-subjects cells subject to the constraint that an equal number appeared in each cell. These samples sizes are similar to those in Rehder (2014a), which tested the same materials and presented related inference questions.

Results

Initial analyses revealed no effects in either experiment of which domain subjects learned, which variable senses were presented, or which version of the data sheet was used, and so the results are presented in Fig. 3a and b collapsed over these factors.

Experiment 1

We first asked whether our manipulation succeeded in changing the overall pattern of ratings shown in Fig. 3a. In fact, a 3 (condition: description-only, description-experience, experience-only) × 7 (inference type: A–H) ANOVA yielded a significant interaction, $F(14, 987) = 5.77$, $MSE = 309.5$, $\eta^2 = .076$, $p < .0001$, confirming that ratings varied depending on whether the causal model was conveyed by verbal description, experience, or both. The pattern of responses in the description-experience condition was significantly different

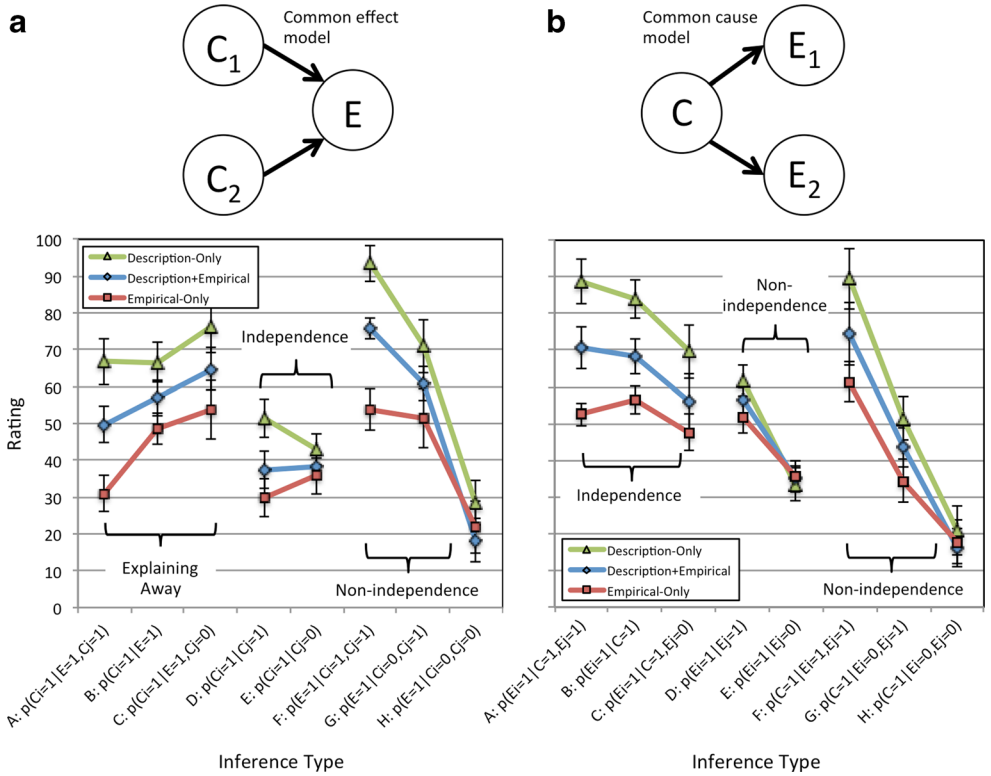

Fig. 3 (**a**) Results from Experiment 1 (common effect condition). (**b**) Results from Experiment 2 (common cause condition). *Error bars* are 95 % confidence intervals

from those in the experience condition, $F(7, 658) = 4.70$, $MSE = 333.9$, $\eta^2 = .048$, $p < .0001$, and marginally different from those in the description condition, $F(7, 658) = 1.64$, $MSE = 274.4$, $\eta^2 = .017$, $p = .120$. We separately present the sets of trials that correspond to explaining away (inference types A, B, and C), independence (D and E), and the nonindependent trials (F, G, and H).

Explaining away (inferences A–C) Explaining away occurs when the probability of a cause given its effect is lower when an alternative cause is present (A < B) and higher when the alternative is absent (B < C). Figure 3a reveals that all three conditions exhibited overall explaining away in that the ratings for inference A were lower than those for C. A 3 (condition) × 3 (inference type) ANOVA revealed a main effect of condition, $F(2, 141) = 27.08$, $MSE = 863.3$, $\eta^2 = .278$, $p < .0001$, of inference type, $F(2, 282) = 42.23$, $MSE = 212.9$, $\eta^2 = .230$, $p < .0001$, and an interaction, $F(4, 282) = 4.91$, $MSE = 212.9$, $\eta^2 = .065$, $p < .001$. A test in which condition (experience-only vs. description-experience vs. description-only) and inference type (A vs. B vs. C) were coded as linear factors also yielded an interaction, $F(1, 94) = 7.19$, $MSE = 285.1$, $\eta^2 = .071$, $p = .001$. The 95 % confidence interval on the difference in the ABC slopes derived from this analysis was [3.3, 22.9], confirming the presence of the negative shift implied by the rich-get-richer principle. Though less positive than in the other conditions, the ABC slope in the description-only condition was significantly greater than zero, $t(47) = 2.53$, $MSE = 348.8$, $\eta^2 = .120$, $p = .015$, reflecting overall explaining away. Nevertheless, Fig. 3a indicates that although this group exhibited normative explaining away on inference types B and C (i.e., B < C), they failed to do so on A and B (A ≅ B).

Independence (inferences D and E) The normative model stipulates that causes are independent in a common effect model and so predicts no difference between inferences D, $p(C_i = 1|C_j = 1)$, and E, $p(C_i = 1|C_j = 0)$. Figure 3a reveals that D < E in the experience-only and description-experience conditions, whereas D > E in the description-only condition. A 3 × 2 ANOVA yielded a main effect of condition, $F(2, 141) = 12.71$, $MSE = 404.4$, $\eta^2 = .153$, $p < .0001$, no effect of inference type, $F < 1$, and an interaction, $F(2, 141) = 7.83$, $\eta^2 = .100$, $p < .001$. Treating condition as a linear factor yielded an interaction, $F(1, 94) = 12.21$, $MSE = 207.5$, $\eta^2 = .115$, $p < .001$, confirming the presence of a negative shift in the DE slope as a causal model was introduced (the 95 % CI on the difference in the slopes was [5.5, 23.6]). That slope was marginally positive in the experience-only condition, $t(47) = 1.98$, $p = .053$, virtually zero in the description-experience condition, $t < 1$, and significantly negative in the description-only condition, $t(47) = 3.00$, $p = .004$.

Nonindependence (inferences F–H) Not surprisingly, subjects judged that the common effect *E* was more likely as the number of causes present increased. As predicted, though, the FGH slope was more negative when a causal model was instructed. A 3 × 3 ANOVA revealed a main effect of condition, $F(2, 141) = 15.75$, $MSE = 1110.7$, $\eta^2 = .183$, $p < .0001$, of inference type, $F(2, 282) = 233.63$, $MSE = 440.9$, $\eta^2 = .624$, $p < .0001$, and an interaction, $F(4, 282) = 8.20$, $MSE = 440.9$, $\eta^2 = .104$, $p < .0001$. Treating condition and inference type as linear factors yielded an interaction, $F(1, 142) = 19.61$, $MSE = 658.8$, $\eta^2 = .173$, $p < .0001$, 95 % CI [18.1, 47.6], confirming the negative shift in the FGH slope.

Experiment 2

The effectiveness of the manipulation was again confirmed by the interaction yielded by a 3 (condition) × 7 (inference type, A–H) ANOVA, $F(14, 987) = 8.21$, $MSE = 264.4$, $\eta^2 = .104$, $p < .0001$. The pattern of responses in the description-experience condition was significantly different from those in both the experience condition, $F(7, 658) = 4.11$, $MSE = 267.1$, $\eta^2 = .042$, $p < .001$, and the description condition, $F(7, 658) = 3.93$, $MSE = 290.1$, $\eta^2 = .040$, $p < .001$. We separately discuss trials corresponding to screening off (inferences A, B and C), and the two nonindependent trial sets (D–E and F–H).

Screening off (inferences A–C) Screening off stipulates that two effects are independent given their common cause, that is, it should be the case that inference types A, B, and C receive the same ratings. Figure 3b reveals subjects in those conditions that were provided with a causal model judged that A > B > C instead, that is, they violated the Markov condition. A 3 (condition) × 3 (inference type) ANOVA revealed a main effect of condition, $F(2, 141) = 34.83$, $MSE = 850.1$, $\eta^2 = .331$, $p < .001$, of inference type, $F(2, 282) = 46.04$, $MSE = 162.9$, $\eta^2 = .246$, $p < .001$, and an interaction, $F(4, 282) = 4.15$, $MSE = 162.9$, $\eta^2 = .056$, $p = .003$. A test in which condition and inference type were coded as linear factors revealed an interaction, $F(1, 94) = 15.89$, $MSE = 158.3$, $\eta^2 = .144$, $p < .001$, 95 % CI [7.0, 21.9]. The ABC slope was negative in all three conditions, ps < .007.

Nonindependence (inferences D and E) The normative model predicts that the two effects are conditionally dependent in the absence of knowledge about the common cause, and Fig. 3b reveals that subjects agreed. Yet the difference between inference D, $p(E_i = 1|E_j = 1)$, and E, $p(E_i = 1|E_j = 0)$, was greater in the conditions that provided a causal model, consistent with the rich-get-richer principle. A 3 × 2 ANOVA yielded no effect of condition, $F(2, 141) = 1.31$, $MSE = 268.7$, $\eta^2 = .018$, $p > .250$, an effect of inference type, $F(1, 141) = 218.58$, $MSE = 157.9$, $\eta^2 = .608$, $p < .001$, and an interaction, $F(2, 141) = 5.87$, $\eta^2 = .077$, $p = .004$. Treating condition as a linear factor yielded an interaction, $F(1, 94) = 13.85$, $MSE = 132.0$, $\eta^2 = .128$, $p < .001$, 95 % CI [5.2, 19.5], reflecting the negative shift in the DE slope as a causal model was introduced.

Nonindependence (inferences F–H) Experiment 1 found that a common effect was rated as more likely when more of its causes were present and Experiment 2 found the analogous result: The common cause was rated as more likely when more of its effects were present, that is, F > G > H. As predicted, though, the FGH slope was more negative when a causal model was instructed. A 3 × 3 ANOVA revealed a main effect of condition, $F(2, 141) = 12.49$, $MSE = 760.2$, $\eta^2 = .151$, $p < .0001$, of inference type, $F(2, 282) = 297.65$, $MSE = 392.8$, $\eta^2 = .679$, $p < .0001$, and an interaction, $F(4, 282) = 4.94$, $MSE = 392.8$, $\eta^2 = .065$, $p < .001$. Treating condition and inference type as linear factors yielded an interaction, $F(1, 94) = 15.45$, $MSE = 485.5$, $\eta^2 = .141$, $p < .001$, 95 % CI [11.2, 38.8], confirming the negative shift in the FGH slope.

Discussion

We found stronger violations of the Markov constraint in the described than the experienced conditions. Subjects failed to fully understand both explaining away (Experiment 1) and screening off (Experiment 2), especially in the conditions that highlighted the causal model. By contrast, these violations were weaker in conditions in which subjects were presented with learning data that allowed them to read off the conditional probabilities. One deviation from our predictions are the slight Markov violations we found in the experience-only conditions (DE in Fig. 3a and ABC in Fig. 3b). We suspect that some subjects misunderstood the inference question, responding with conjunctive (i.e., $p(X, Y)$) rather than conditional probabilities ($p(X|Y)$).

General discussion

Research in the past two decades has shown that causal Bayes nets capture many aspects of human causal thinking—such as sensitivity to differences in causal direction and between inferences based on interventions versus observations—that sets this type of reasoning apart from purely associative or logical reasoning. However, central properties of this normative theory are routinely violated. This research focused on failures of explaining away and screening off by comparing conditions in which causal models were merely described versus experienced. Using this design, we followed the lead of the judgment and decision-making literature, which has revealed interesting dissociations in these two tasks, showing that biases previously considered universal often depend on how scenarios are presented (Hertwig, 2015).

Our key finding is that we obtained stronger deviations from the normative causal Bayes net model of causal reasoning in the conditions that described causal models compared to those that presented learning data. The finding in the described conditions is consistent with previous studies demonstrating an associative bias in causal reasoning (Rehder, 2014a). We found that subjects' inferences conformed better to the normative reasoning implied by causal Bayes nets in a condition that is typically used to study associative learning as compared to one that presented a causal model. Although we found slight biases in the experience-only condition as well (which may be due to misunderstandings of the test questions), subjects were overall relatively competent in estimating probabilities from data.

Perhaps the most interesting comparison is between the experience-only and the description-experience conditions because in both subjects can access conditional probabilities directly from the provided data. Recall that because there is some flexibility in how test questions are translated into probability inferences, the experience-only condition serves as an important control for how subjects understand the test questions. Here we found that adding information about the underlying causal model in the description-experience condition led to *stronger* violations of the normative implications of Bayes nets, supporting our claim that causal models induce a rich-get-richer bias. Because it conveyed qualitative causal model information but no statistics, judgments in the description-only condition are less comparable to those in the other two conditions in an absolute sense. Nevertheless, it is meaningful to compare relative judgments across conditions (i.e., to compare the slopes of the six lines in Fig. 3), and in fact we found, as predicted, that the influence of the rich-get-richer principle was strongest in the description-only condition.

It is important to note that the rich-get-richer principle does not imply that reasoners ignore the direction of causality in their causal inferences. Common cause and common effect networks are theoretically important because they are structurally identical, ignoring the direction of causality. Yet in every experimental test of which we are aware (including this one), these networks elicit different inferences (compare Fig. 3a and b). The conclusion to be drawn from this work is not that people aren't sophisticated causal reasoners but rather that their inferences are a product of an interaction between the normative model and the rich-get-richer principle. Consistent with this interpretation, a computational model that is under development and that combines the rich-get-richer principle with predictions derived from causal Bayes can successfully reproduce the present results (Rehder, 2016). In particular, it simultaneously accounts for failures of both explaining away and screening off.

Past explanations of the rich-get-richer principle

Having established the existence of the rich-get-richer effect in causal reasoning, it is natural to ask why it occurs. As mentioned in the introduction, a number of alternative accounts have been proposed. But although it is likely that each of these accounts contribute to the failures of explaining away and screening off in specific cases, none provide a comprehensive account of the failures observed here and in the literature more broadly.

One class of explanation posits that the specific prior knowledge that subjects bring to the experimental situation influences the causal model representation people reason with. For example, in the current experiments subjects who were taught that low interest rates, large trade deficits, and high retirement savings were causally related might have had preexisting beliefs about how those variables were causally related, beliefs that just so happened to yield (apparent) Markov violations and weak explaining away. Yet this conjecture fails to account for the results from the present experiments because of the use of counterbalanced materials. Because some subjects were told that low interest rates cause high retirement savings and others that low interest rates cause *low* retirement savings (and still others were told that *high* interest rates cause high retirement savings, etc.). That is, any prior knowledge that subjects might have possessed about interest rates and retirement savings canceled out by averaging over subjects.

Some accounts that appeal to specific domain knowledge only pertain to certain causal network topologies. Park and Sloman (2013, 2014) demonstrated that reasoners instructed on a common cause model and then confronted with a test question that specifies, as part the premise, a situation in which the cause C is present and an effect (say E_1) is absent will augment the model with a disabler to explain why C didn't produce E_1. Moreover, if they view $C \rightarrow E_1$ and $C \rightarrow E_2$ as sharing underlying causal mechanism then the factor that potentially disables $C \rightarrow E_1$ may, when present, also disable $C \rightarrow E_2$, (with the result that the absence of E_1 is evidence for the presence of the disabler which in turn is evidence for the absence of E_2, that is, an [apparent] violation of screening off occurs). Yet not only does this account fail to explain the screening off errors that arise when $C \rightarrow E_1$ and $C \rightarrow E_2$ do *not* share a mechanism (as they didn't in the current Experiment 2), it provides no account of the errors that arise with a common effect network (unconditionally dependent causes and weak explaining away).

Perhaps the strongest evidence against accounts based on domain-specific knowledge comes from independence violations observed even when no domain is specified. For example, Rehder (2014a) found that an abstract condition in which the variable names were simply letters (*A, B*, and *C*) yielded independence violations as large as those that used meaningful

content (e.g., interest rates, trade deficits, and retirement savings).

Other approaches appeal not to domain knowledge per se but rather to the abstract expectations that reasoners have about the reasoning situation. Inspired by dispositional theories of causation (e.g., Wolff, 2007), Mayrhofer and Waldmann (2015) proposed that people distinguish between causal agents and causal patients participating in causal relations (e.g., in "Wind moves the boat," wind is the causal agent and the boat the patient). In most real-world causal models, the causes are associated with the agent role. Mayrhofer and Waldmann's studies showed that in situations in which an agent fails to produce the expected effect, a more general failure of the capacity of the agent is inferred. Thus, the violations of screening off that obtain with a common cause network are expected when the common cause is represented as the causal agent. A novel prediction, which was empirically confirmed, is that Markov violations should be diminished in the rare cases in which causal agents are associated with the effect events. This theory explains the Markov violations in abstract scenarios under the assumption that agents are per default associated with the cause event (Waldmann & Mayrhofer, 2016). Just as with Park and Sloman (2013, 2014), however, it provides no account of the errors associated with common effect networks.

Rehder and Burnett (2005) appealed to abstract expectations to explain the Markov violations they observed with variables that were causally related features of categories (also see Rehder, 2014b). They proposed that people believe that many categories possess underlying causal processes that bring rise to observed features, a view related to the well-known *essentialist* intuitions about categories (Gelman, 2003; Medin & Ortony, 1989). Such a view reproduces the rich-get-richer effect: The presence of one feature implies the presence of normally operating causal processes that in turn imply the presence of other features. But although this account doesn't assume specific domain knowledge, it is specific to features of categories and so doesn't apply to the materials tested in most studies of causal reasoning errors (including this one).

Nevertheless, it is tempting to try to extend this account to non-categorical materials. After all, it is reasonable to suppose that subjects in the current experiments assumed that the economic, meteorological, and sociological variables were related in ways other than those we specified, even if they had no specific ideas about what those relationships were. However, recognize that the assumption that variables are related doesn't explain the *direction* of the observed violations. Merely believing that, say, interest rates, trade deficits, and retirement savings are somehow causally related doesn't explain why, with a common cause network, $p(E_i = 1|C = 1,\ E_j = 1)$ was greater rather than less than $p(E_i = 1|C = 1, E_j = 0)$ or why, with a common effect network, $p(C_i = 1|C_j = 1)$ was greater rather than less than $p(C_i = 1|C_j = 0)$ (or why explaining away was too weak rather than too strong).

Additional assumptions are thus required to explain why the variables might be causally related in just the way required to yield the observed inferences. Rehder (2014a) raised the possibility that a search of memory triggered by comprehending the materials may be biased toward other generative relations involving the causally related variable senses (e.g., reading that "low interest rates cause small trade deficits" may yield facts about how low interest rates and small trade deficits are positively related, but not facts about how they are negatively related, or how *high* interest rates and small trade deficits are positively related). Or, perhaps reasoners spontaneously *construct* other ways of how low interest rates and small trade deficits are generatively related. Then, perhaps their experience with retrieving or constructing causal models with many generative relations generalizes to abstract reasoning scenarios. A very different sort of approach would be to imagine that the theories that scientists have discovered (or at least the ones that people are taught) include variables that tend to be generatively related in multiple ways. People then generalize this fact to all domains about which they reason (including abstract ones).

In summary then, none of the proposed accounts provide a full satisfactory account of the documented reasoning errors. Although accounts based on abstract expectations are clearly more promising than ones based on specific domain knowledge, they are so far either restricted to certain network topologies (common cause networks) or require additional assumptions that will remain speculative until additional evidence in their favor is established (e.g., biased memory search). Further research will be required to determine if, for example, abstract accounts can be developed for other network topologies (such as common effect networks). Alternatively, the rich-get-richer principle may prove to be a ubiquitous feature of human causal reasoning, one that manifests itself consistently across content domains, abstract expectations, and network topologies.

Perspectives on future research

There are a number of open questions regarding the nature of the rich-get-richer principle. In the introduction we noted that our use of "association"—bidirectional links between events—differs from its use in in traditional associative learning theory in which cue competition between redundant cues plays an important role (e.g., Rescorla & Wagner, 1972). One open question concerns how a parameter that represents the magnitude of the rich-get-richer principle is estimated from learning data along with the

rest of the parameters of a causal model (Lagnado et al., 2007; Rottman, in press; Waldmann, 1996). One possibility raised by some researchers is that people may use two different learning mechanisms, an associative and a causal one, with context and performance factors determining which of the two is active (see, e.g., Jocham et al., 2016; Le Pelley, Griffiths, & Beesley, in press). Although this model, given its focus on learning, is very different from ours, it would be interesting to study whether there are interrelations between the two.

Alternative interpretations of the rich-get-richer principle exist. One reason that reasoners might violate independence is that activation "spreads" along the nodes of a causal net in much the same it is thought to do so in memory, even in cases when the flow of information is normatively (according to causal Bayes nets) blocked. For example, in a common cause model in which the state of the common cause is known, information about one effect might "leak" through the common cause node to the other effect (violating independence). An alternative interpretation is that reasoners conceive of causal models as having "prototypical" states and, in a particular reasoning situation, compute the similarity between that situation and the most similar prototypical state—and predict that the to-be-predicted variable will take on the value in the prototype as a function of that degree of similarity. For example, a situation in which a common cause and an effect are both present is more similar to a prototype network state (all variables present) than one in which only the common cause is present; thus, the other effect is more likely to be present in the former (violating independence). Although the present results do not allow us to distinguish these possibilities, Rehder (2016) presented causal models that were more complicated than those tested here and found that subjects' inferences were more consistent with reasoning with respect to prototypical network states. This result naturally leads to the question of how causal inferences are affected by the frequency of prototypical states in learning data.

Finally, the vast majority of studies of causal reasoning failures have investigated models with independent, generative links. Little research has studied independence failures with alternative functional forms (e.g., when causes combine conjunctively rather than independently; although see Rehder, 2014b) or when causal relations are preventative rather than generative. Open questions include, for example, what "prototypical" network states are implied by a causal model with a mixture of generative and preventative causal relations. In such cases the prototype might embody negative rather than positive associations among some variables.

In summary, our findings indicate that neither a noncausal associative nor a purely normative causal theory, such as Bayes nets, fully captures human causal reasoning. An integrated formal theory that integrates nonnormative biases with normative sensitivity to causal features is an important goal for future research.

Author note Bob Rehder, Department of Psychology, New York University, USA. Michael R. Waldmann, Department of Psychology, University of Göttingen, Germany.

Appendix

Materials

Table 2 presents the three variables used in the domains of economics, meteorology, and sociology. Recall from the main text that the variables were described as binary to subjects (e.g., interest rates were either low or normal). To control for any domain knowledge that subjects might have brought to the experiment, a between-subjects factor that took on the values +++, +--, -+-, and --+ controlled which variable states were used, where each +/- picks out the value in Table 2. For example, the nonnormal values for interest rates, trade deficits, and retirement savings were either (low, small, high), (low, large, low), (high, small, low), or (high, large, high) in the +++, +--, -+-, and --+ conditions, respectively.

Table 2 Variables

Economics	Meteorology	Sociology
Interest rates	Ozone levels	Urbanization
(low+/high–)	(high+/low–)	(high+/low–)
Trade deficits	Air pressure	Interest in religion
(small+/large–)	(low+/high–)	(low+/high–)
Retirement savings	Humidity	Socio-economic mobility
(high+/low–)	(high+/low–)	(high+/low–)

Table 3 presents examples of the causal relationships in the domain of economics in the +++ counterbalancing condition. The different variable values in the four counterbalancing conditions required different causal relationships of course. For instance, the relationship between interest rates and trade deficits was described as low → small, low → large, high → small, and high → large in the +++, +--, -+-, and --+ conditions, respectively. The counterbalancing thus required the use of 12 distinct causal relationships in each of the three domains. See Appendix A of Rehder (2014a) for all the causal relationships.

Table 3 Example causal relationships in the domain of economics

Common effect causal relationships	Causal mechanism
Low interest rates → High retirement savings	Low interest rates cause high retirement savings. Low interest rates stimulate economic growth, leading to greater prosperity overall, and allowing more money to be saved for retirement in particular.
Small trade deficits → High retirement savings	Small trade deficits cause high retirement savings. When the economy is good, people can cover their basic expenses and so have enough money left over to contribute to their retirement accounts.
Common cause causal relationships	**Causal mechanism**
Low interest rates → Small trade deficits	Low interest rates cause small trade deficits. The low cost of borrowing money leads businesses to invest in the latest manufacturing technologies, and the resulting low-cost products are exported around the world.
Low interest rates → High retirement savings	Low interest rates cause high retirement savings. Low interest rates stimulate economic growth, leading to greater prosperity overall, and allowing more money to be saved for retirement in particular.

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology, 35,* 303–314.

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

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44,* 211–233.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General, 140*(2), 168–185.

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

Gelman, S. A. (2003). The essential child: The origins of essentialism in everyday thought. New York: Oxford University Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 3–23.

Griffiths, T. (in press). Formalizing prior knowledge in causal induction. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning.* New York, NY: Oxford University Press.

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

Hagmayer, Y., & Meder, B. (2013). Repeated causal decision making. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 39,* 33–50.

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

Hausman, D. M., & Woodward, J. (1999). Independence, invariance, and the causal Markov condition. *British Journal for the Philosophy of Science, 50,* 521–583.

Hayes, B. K., Hawkins, G. E., Newell, B. R., Pasqualino, M., & Rehder, B. (2014). The role of causal models in multiple judgments under uncertainty. *Cognition, 133,* 611–620.

Hertwig, R. (2015). Decisions from experience. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (Vol. 1, pp. 240–267). Chichester: Wiley Blackwell.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15,* 534–539.

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General, 139,* 702–727.

Jocham, G., Brodersen, K. H., Constantinescu, A. O., Kahn, M. C., Ianni, A. M., Walton, M. E., & Behrens, T. E. J. (2016). Reward-guided learning with and without causal attribution. *Neuron, 90,* 177–190.

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

Kelley, H. H. (1972). *Causal schemata and the attribution process*. New York: General Learning Press.

Kemp, C., Shafto, P., & Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology, 64,* 35–73.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116,* 20–58.

Khemlani, S. S., & Oppenheimer, D. M. (2010). When one model casts doubt on another: A levels-of-analysis approach to causal discounting. *Psychological Bulletin, 137,* 1–16.

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 136,* 430–450.

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

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). New York: Oxford University Press.

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 754–770.

Le Pelley, M., Griffiths, O., & Beesley, T. (in press). Associative accounts of causal cognition. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning*. New York, NY: Oxford University Press.

Lee, H. S., & Holyoak, K. J. (2008). The role of causal models in analogical inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1111–1122.

Lejarraga, T., Pachur, T., Frey, R., & Hertwig, R. (2016). Decisions from experience: From monetary to medical gambles. *Journal of Behavioral Decision Making, 29,* 67–77.

Liljeholm, M., & Cheng, P. W. (2007). When is a cause the "same"? Coherent generalization across contexts. *Psychological Science, 18,* 1014–1021.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology, 61,* 303–332.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115,* 955–982.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science, 34,* 113–147.

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

Mayrhofer, R., & Waldmann, M. R. (in press). Sufficiency and necessity assumptions in causal structure induction. *Cognitive Science.*

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

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review, 121,* 277–301.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 179–196). Cambridge, MA: Cambridge University Press.

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

Oppenheimer, D. M., Tenenbaum, J. B., & Krynski, T. R. (2013). Categorization as causal explanation: Discounting and augmenting in a Bayesian framework. *Psychology of Learning and Motivation, 58,* 203–231.

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

Park, J., & Sloman, S. A. (2014). Causal explanation in the face of contradiction. *Memory & Cognition, 1,* 1–15.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Perales, J., Catena, A., Cándido, A., & Maldonado, A. (in press). Rules of causal judgment: Mapping statistical information onto causal beliefs. In M. R. Waldmann (Ed.), *Oxford handbook of causal reasoning.* New York, NY: Oxford University Press.

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

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

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

Rehder, B. (2006). When causality and similarity compete in category-based property induction. *Memory & Cognition, 34,* 3–16.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science, 33,* 301–343.

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

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

Rehder, B. (2016). Beyond Markov: Accounting for independence violations in causal reasoning. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1853–1858). Austin: Cognitive Science Society.

Rehder, B. (in press-a). Categories as causal models: Categorization. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning*. New York: Oxford University Press.

Rehder, B. (in press-b). Categories as causal models: Induction. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning*. New York: Oxford University Press.

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

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

Rehder, B., & Kim, S. (2009). Classification as diagnostic reasoning. *Memory & Cognition, 37,* 715–729.

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

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science, 34,* 175–221.

Rips, L. J., & Edwards, B. J. (2013). Inference and explanation in counterfactual reasoning. *Cognitive Science, 37,* 1107–1135.

Rottman, B. (in press). The acquisition and use of causal structure learning. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning.* New York, NY: Oxford University Press.

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

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

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition, 109,* 175–192.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 21, pp. 229–261). New York: Academic Press.

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science, 29,* 5–39.

Spirtes, P., Glymour, C., & Scheines, P. (1993). *Causation, prediction, and search*. New York: Springer-Verlag.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27,* 453–489.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale: Erlbaum.

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation, vol. 34: Causal learning* (pp. 47–88). San Diego: Academic Press.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 53–76.

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science, 31,* 233–256.

Waldmann, M. R. (in press). *The Oxford handbook of causal reasoning*. New York: Oxford University Press.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82,* 27–58.

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.

Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed.), *Oxford handbook of cognitive psychology* (pp. 733–752). New York: Oxford University Press.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121,* 222–236.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General, 124,* 181–206.

Waldmann, M. R., & Mayrhofer, R. (2016). Hybrid causal representations. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 65, pp. 85–127). New York: Academic Press.

Walsh, C. R., & Sloman, S. A. (2004). *Revising causal beliefs.* Paper presented at the Proceedings of the 26th Annual Conference of the Cognitive Science Society, Mahwah, NJ.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General, 136,* 82–111.