Argument and Computation
2012, 1–25, iFirst

Taylor & Francis
Taylor & Francis Group

Cognitive shortcuts in causal inference

Philip M. Fernbach[a]* and Bob Rehder[b]

[a]*Leeds School of Business, University of Colorado, 19 UCB, Boulder, CO 80309-0419, USA;* [b] *Department of Psychology, New York University, NY, USA*

(*Received 8 December 2011; final version received 28 March 2012*)

The paper explores the idea that causality-based probability judgments are determined by two competing drives: one towards veridicality and one towards effort reduction. Participants were taught the causal structure of novel categories and asked to make predictive and diagnostic probability judgments about the features of category exemplars. We found that participants violated the predictions of a normative causal Bayesian network model because they ignored relevant variables (Experiments 1–3) and because they failed to integrate over hidden variables (Experiment 2). When the task was made easier by stating whether alternative causes were present or absent as opposed to uncertain, judgments approximated the normative predictions (Experiment 3). We conclude that augmenting the popular causal Bayes net computational framework with cognitive shortcuts that reduce processing demands can provide a more complete account of causal inference.

Keywords: cognitive science < interdisciplinary links with computational argument; conditionals < interdisciplinary links with computational argument; mental models < interdisciplinary links with computational argument; rationality < interdisciplinary links with computational argument; computational accounts of probabilistic argument; explanation

Downloaded by [philip fernbach] at 05:47 07 June 2012

The psychology of causal inference is experiencing growing pains. A proliferation of interest in causal reasoning over the last several years is due in large part to the development of causal Bayesian networks, a computational framework for learning, representing and reasoning with causal knowledge. Causal Bayes nets are normative models that are governed by the axioms of probability, and psychological theories based on causal Bayes nets therefore predict that causal judgment should accord with norms. In line with this idea is a variety of evidence that people (including young children) are sophisticated and adept causal reasoners. However, as we detail below, probability judgments based on causal evidence do not always honor the norms associated with Bayes nets, suggesting that understanding such judgments will require considering nonnormative factors imposed by the cognitive processes that implement causal reasoning.

We will argue that causal judgments can be viewed as emerging from an interaction between two competing drives: one towards veridicality and one towards effort reduction. We will illustrate this claim for a paradigmatic task, namely, judging the conditional probability of a hypothesis given causally relevant evidence. We present an analysis of the requirements for optimal performance on this task, one that suggests a number of task variables that may invite reasoners to reduce effort by taking shortcuts that result in inappropriate conclusions. In three experiments we manipulate these variables with an eye to identifying which normative requirements people typically violate and to establishing conditions that support more veridical judgment.

*Corresponding author. Email: philip.fernbach@gmail.com

ISSN 1946-2166 print/ISSN 1946-2174 online
© 2012 Taylor & Francis
http://dx.doi.org/10.1080/19462166.2012.682655
http://www.tandfonline.com

Downloaded by [philip fernbach] at 05:47 07 June 2012

Causal Bayes nets and the normativity of causal inference

People are good at qualitative, contextualised reasoning about the causal systems they interact with in their lives (Sloman 2005). People often have good intuitions for instance about what actions to take to achieve a goal, what caused them to feel ill, or how a new coach will influence the performance of their favorite team. This capability has prompted substantial interest in a theory of mental representation that accounts for causal intuitions: a causal model theory (Waldmann and Holyoak 1992; Glymour 1998; Gopnik, Glymour, Sobel, Schulz and Kushnir 2004). Causal model theories are usually instantiated using causal Bayesian networks, graphs where events or properties and their causal relations are depicted as variable nodes and directed edges (arrows) that point from cause to effect (Spirtes, Glymour and Scheines 1993; Pearl 1988, 2000; Jordan 1999). A causal Bayes net is associated with functions that specify how the probabilities of effects change in the presence of their causes. These functions allow for the calculation of the probability of unknown variables conditioned on known ones and thus support inductive inference.

Since causal Bayes nets are based on probabilistic calculus, psychological theories based on these models predict that human causal inference should respect probabilistic norms when people can apply an appropriate causal model. A variety of evidence supports this idea: People honour many causal reasoning norms not only during simple inferences (Rehder and Burnett 2005) but also more complex causal inferences involving analogies (Lee and Holyoak 2008; Holyoak, Lee and Lu 2010), generalisations (Rehder and Hastie 2004; Rehder 2006, 2009; Shafto, Kemp, Bonawitz, Coley and Tenebaum 2008; Kemp and Tenenbaum 2009), and acts of classification (Rehder and Hastie 2001; Rehder 2003a, b; Rehder and Kim 2006, 2009, 2010). Research on development shows that even children's reasoning is more consistent with causal than associative models by age 4 (Sobel, Tenenbaum and Gopnik 2004; Hayes and Thompson 2007; Opfer and Bulloch 2007; Hayes and Rehder in press). Finally, Causal Bayes nets are also attractive to learning theorists because causal structures and parameters can (in principle) be learned from data (Waldmann et al., 1995; Cheng 1997; Novick and Cheng 2004; Griffiths and Tenenbaum 2005, 2009; Lu, Yuille, Liljeholm, Cheng and Holyoak 2008;, but see Fernbach and Sloman 2009). Causal models provide a better account than associative models for such learning in both adults (Waldmann and Holyoak 1992; Waldmann 2000; for a review see Holyoak and Cheng 2011) and children (Gopnik et al. 2004; Sobel et al. 2004), and sometimes nonhumans (Beckers, De Houwer, Miller and Urushihara 2006; Blaisdell, Sawa, Leising and Waldmann 2006). In summary, causal Bayes nets have provided a valuable organising framework for a large variety of reasoning and learning phenomena.

Despite these successes, other phenomena suggest that causal inference is error-prone. Many counternormative phenomena from the heuristics and biases literature – such as conjunction fallacies (Tversky and Kahneman 1983), subadditive probability judgments (Tversky and Koehler 1994; Rottenstreich and Tversky 1997), simulation effects (Kahneman and Tversky 1982; Wells and Gavanski 1989), and hindsight biases (Fischhoff and Beyth 1975) emerge (and are sometimes exacerbated) in causal scenarios. Tversky and Kahneman (1980) argue that causal reasoning is qualitatively different from a more appropriate evaluation of evidential strength and therefore leads to biased judgment. Moreover, people sometimes confuse the causal role of their actions. This leads to 'diagnostic self-deception' (Quattrone and Tversky 1984; Sloman, Fernbach and Hagmayer 2010) and other examples of 'evidential reasoning' such as cooperation in the prisoner's dilemma and the voter's illusion (Acevedo and Krueger 2005). People also sometimes feel a false sense of control over outcomes that are actually up to chance or risk (Langer 1975) leading to idiosyncratic superstitions like reluctance to "tempt fate" (Risen and Gilovich 2008; Swirsky, Fernbach and Sloman 2011). In the causal learning literature as well, researchers have documented conditions in which learners depart from the normative acquisition rules specified by Bayes nets (De Houwer and Beckers, 2003; Reips and Waldmann 2008; Waldmann and Walker 2005).

Yet a different kind of error was uncovered in a series of studies by Fernbach, Darlow and Sloman (2010, 2011a). Following Tversky and Kahneman (1980) they compared predictive reasoning – judgment of the conditional probability of an effect given a cause – to diagnostic reasoning – judgment of the conditional probability of a cause given an effect. By varying causal structure and collecting judgments of conditional probability about a variety of scenarios, they were able to evaluate the consistency of judgments in both directions of reasoning. In predicting an effect from a cause, participants systematically neglected the contribution of alternative causes to the probability of the effect. They based their judgments just on the strength of the cause known to be present, and therefore gave conditional probability judgments that were too low. In contrast, diagnostic judgments were sensitive to the strength of alternative causes and approximately consistent with the predictions of a causal Bayes net model.

Fernbach, Darlow and Sloman (2011b) established the robustness of the neglect of alternative causes in a series of experiments assessing conditional probability judgments and gambling decisions in the face of a weak but positive predictive evidence. Ignoring alternative causes can be a serious error when the conditional probability is high, but the contribution of the given cause to that probability is small. Indeed, Fernbach et al. found cases where the conditional probability of the effect given a weak cause is judged lower than the marginal probability of an effect (i.e. the probability of the effect when no evidence is mentioned). For instance, participants told about weak but positive evidence that the Republicans would win the House of Representatives in the 2010 US mid-term election (a newspaper endorsement of a single candidate) were actually less likely to gamble on the Republicans winning than participants given no evidence. Apparently, the focus on the single cause mentioned in a conditional judgment crowded out other causes that would otherwise be considered. Fernbach et al. refer to this as the *weak evidence effect*.

Downloaded by [philip fernbach] at 05:47 07 June 2012

Effort reduction as a reconciliatory principle

Why might people violate the norms of causal Bayes nets and probability theory more generally? A few authors have tried to bring theoretical organisation to the heuristics and biases literature by appealing to *effort reduction* as a fundamental drive of human cognition. Kahneman and Frederick (2002) argue that a heuristic is a surreptitious substitution of an easy question for a hard one. More recently, Shah and Oppenheimer (2008) taxonomised a large number of heuristics according to their role in reducing effort relative to what would be required by a full optimal solution to the weighted-additive choice rule (Payne, Bettman and Johnson 1993). They argue that the optimal solution is out of reach because it requires a complex series of processes with many inputs and computational demands. One might argue that the appeal to effort reduction is too vague to provide much explanatory power on its own. We agree with this point, and our goal in this paper is not to litigate this issue but merely to demonstrate the types of shortcuts people make in causal reasoning.

Like the weighted-additive rule, normative causal inference imposes substantial computational demands. Consider the many steps required to render a causal-based judgment of conditional probability: first, a qualitative representation or model of the causal situation must be constructed. This involves not only identifying the causal relations that directly relate evidence and hypotheses but also filling in additional causal variables that may be relevant to the judgment, including alternative causes, enabling conditions, disabling conditions, and so forth. At this step, two sorts of errors may arise. *Errors of omission* may occur when relevant variables are not included in the model. This may occur because a reasoner's cursory search of long-term memory may fail to yield all relevant knowledge. In contrast, *errors of commission* occur when relevant knowledge that is readily available (e.g. already retrieved from memory, supplied as part of the reasoning problem, etc.) is nevertheless ignored by the reasoner.

Next, one must identify the functional relations by which causes bring about effects and parameterise those relations (e.g. with the strength of the causal relations). Judgment errors may arise at this stage if causal relations are represented in a simplified form (e.g. as a symmetric associative relation, Rehder 2009) or if causal strengths are represented with low fidelity (e.g. qualitatively rather than quantitatively).

Third, to assess the net influence of hidden variables, the reasoner must integrate over their possible states. To illustrate the importance of integrating over the states of potential alternative causes, it is useful to consider the normative equations for causal inferences based on a reasonably general noisy or parameterisation associated with generative causes between binary variables (for details see Waldmann, Cheng, Hagmayer and Blaisdell 2008; Rehder 2010; Fernbach et al. 2011a). Equation (1) specifies the probability of the effect given the cause assuming that it can be brought about by the focal cause itself (with probability W_{C}) *or* by one or more alternative causes (with probability W_{NetAlt}). Equation (2) shows that the probability of the cause given the effect is also a function of W_{C} and W_{NetAlt} (in addition to the base rate of the cause, P_{C}). Finally, Equation (3) specifies how W_{NetAlt} summarises the net effect of Nalternative (independent and generative) causes. Importantly, the effect of an alternative cause A_i depends on the strength of the causal relation linking it with the effect (W_{A_i}) times the probability that A_i is in fact present (P_{A_i}). Note that when there is only a single alternative cause A, Equation (3) reduces to $P_A W_A$

$$P(\text{Effect}|\text{Cause}) = W_{\mathrm{C}} + W_{\mathrm{NetAlt}} - W_{\mathrm{C}} W_{\mathrm{NetAlt}}, \tag{1}$$

$$P(\text{Cause}|\text{Effect}) = 1 - (1 - P_{\mathrm{C}}) \frac{W_{\mathrm{NetAlt}}}{P_{\mathrm{C}} W_{\mathrm{C}} + W_{\mathrm{NetAlt}} - P_{\mathrm{C}} W_{\mathrm{C}} W_{\mathrm{NetAlt}}}, \tag{2}$$

$$W_{\mathrm{NetAlt}} = 1 - \prod_{i=1,\ldots,N} (1 - W_{A_i} P_{A_i}). \tag{3}$$

Equation (3) suggests a number of shortcuts that reasoners may take in accounting for the influence of alternative causes. For example, for each alternative cause A_i reasoners may assume values for P_{A_i} and W_{A_i}that reduce the effort involved in computing $P_{A_i} W_{A_i}$: assume the alternative is always present ($P_{A_i} = 1$), always absent ($P_{A_i} = 0$), or always effective ($W_{A_i} = 1$). Each of these three possibilities implies a particular type of judgment error: alternative causes are ignored entirely in the second, their strength is ignored in the third, and their influence is overestimated in the first.

Fourth, once a causal model is constructed, parameterised, and the effect of hidden causal factors computed, reasoners must *aggregate* the influence of the focal and alternative causes to render a judgment of conditional probability. Errors may be introduced at this stage if they choose to use a qualitative combination rule instead of Equation (1) or (2).

Overview of experiments

The analysis just presented suggests a number of hypotheses regarding why errors arise during causal inferences: (a) relevant variables may not be retrieved from memory, (b) their representation may be deleted from the causal model, (c) integration may rely on shortcuts that misrepresent their influence, and (d) alternatives may not be aggregated appropriately with the focal cause. The aim of the following experiments was to conduct a first assessment of the extent to which each of these shortcuts influence causal reasoning. We taught people novel categories by describing a category's features, its causal model (the structure of its interfeature causal relations), and the model's parameters (the strengths of those relations). After being trained on a novel category, subjects were presented with a category member that possessed one or more features and asked for

Downloaded by [philip fernbach] at 05:47 07 June 2012

the likelihood that another feature was present. Following Fernbach et al. (2011a), we varied both the direction of inference (i.e. predicting effects from causes and diagnosing causes from effects) and the strengths of the focal and alternative causes. Specifying causal strengths also allowed us to calculate the normative responses to the inference questions and thus identify conditions that lead to errors. In Experiment 1, we extended Fernbach et al.'s (2011a) design to a task with novel categories where memory retrieval requirements were minimised (participants were provided with a diagram of the category's causal model during the inference task). In Experiment 2, we varied the number and explicitness of alternative causes and the computational difficulty of aggregating parameters. In Experiment 3, we varied whether alternative causes are unknown, known to be present or known to be absent in a particular category member.

Experiment 1

Participants in Experiment 1 were taught one of the two category structures in Figure 1. All subjects learned categories with four binary features. Feature C_w was described as the cause of E_w and C_s was described as the cause of E_s. Examples of features and causal relationships are shown in Table 1. Our central manipulation concerned the different strengths of the alternative causes of the two effect features. To convey the presence of alternative causes, both E_w and E_s were described as also being caused by "one or more" unnamed category features. However, the alternative causes of E_w were described as relatively weak (hence the "w" subscript) by stating that

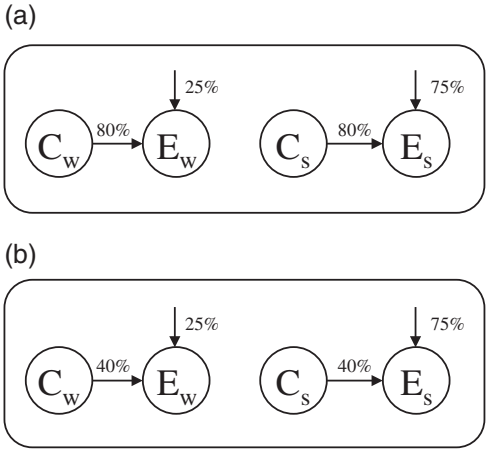

Figure 1. Causal structure tested in Experiment 1: (A) strong focal cause condition and (B) weak focal cause condition.

Table 1. Features and causal relationships for Myastars, an artificial category.

Feature	Causal relationship
Ionised helium	Ionised helium causes the star to be very hot. Ionised helium participates in nuclear reactions that release more energy than the nuclear reactions of normal hydrogen-based stars, and the star is hotter as a result
Very hot temperature	
High density	High density causes the star to have a large number of planets. Helium, which cannot be compressed into a small area, is spun off the star, and serves as the raw material for many planets
Large number of planets	

Downloaded by [philip fernbach] at 05:47 07 June 2012

E_w appeared in category members with probability 25% even when C_w was absent. In contrast, the alternative causes of E_s were described as relatively strong ("s") by stating that it appeared in category members with probability 75% even when C_s was absent.

After learning, participants were asked a series of conditional probability questions. On predictive questions they were told that a particular category member possessed a cause feature and were asked to judge the likelihood that it possessed the relevant effect feature, and vice versa for diagnostic questions. Subjects were also asked to predict the cause and effect given the *absence* of the effect and cause, respectively. Finally, subjects were asked to make unconditional (i.e. marginal) judgments by estimating the prevalence of each feature in a category. Effects should be judged to be more prevalent to the extent they have strong *vs*. weak alternative causes.

Unlike in Fernbach et al.'s (2011a) studies, participants did not have to retrieve causal knowledge from memory. Participants learned the novel interfeature causal relations as part of the experiment and were provided with a diagram of the causal relations during the inference test. Participants were also provided with explicit information regarding both the functional form of the relationship (in the form of a description of the causal mechanism by which the cause generates the effect) between the cause and effect features and the strength of those relationships (by specifying how often a cause, when present, would generate its effect). Finally, the effect of alternative causes was provided in summary form, eliminating the need to integrate over hidden causes. That is, by telling them it was either weak (25%) or strong (75%), reasoners were directly provided with the value of W_{NetAlt} (the net effect of all alternative causes), relieving them of the need to compute it via Equation (3).

A secondary objective of Experiment 1 was to assess whether the neglect of alternative causes depends on the strength of the focal cause itself. To this end, the strengths of the focal causes (between C_w and E_w, and C_s and E_s) were manipulated as a between-subjects variable. Half of the subjects were told that the strength of these relationships was 40% (Figure 1(A)), whereas the other half was told that their strength was 80% (Figure 1(B)). This manipulation is of theoretical interest because it speaks about the possibility that reasoners exhibit *strategic laziness*, they neglect alternative causes only when doing so is unlikely to yield large errors in judgments. When a focal cause is strong (e.g. 80%), the maximum error in predictive inference cannot exceed 20% (because the effect cannot be more probable than 100%), whereas it may be as large as 60% for a weak focal cause of 40%. This raises the possibility that subjects in Experiment 1 may be less likely to neglect alternative causes in the weak focal cause condition than the strong one.

We summarise by presenting the normative predictions for this experiment in Figure 2.[1] First, normative predictions for predictive inferences (computed from Equation (1)) are shown in Figure 2(A). This panel confirms that such inferences ought to be sensitive to alternative cause strength and this sensitivity ought to be larger for weak (40%) *vs*. strong (80%) focal causes. Second, Figure 2(B) shows that diagnostic inferences (Equation (2)) should also be sensitive to alternative causes and, on the basis of previous research, we predict that they will be in this experiment. The probability of an effect given the absence of the focal cause (panel C) is simply the net strength of the alternatives (W_{NetAlt}). Finally, Figure 2(D) reveals that predicting a cause given the absence of the effect (panel D) should be sensitive to the strength of the focal cause but not the strength of the alternative cause.[2] We assess where the inferences made by human causal reasoners diverge from those in Figure 2.

Method

Materials

Six novel categories were tested: two biological kinds (Kehoe Ants and Lake Victoria Shrimp), two nonliving natural kinds (Myastars [a type of star] and Meteoric Sodium Carbonate), and

Downloaded by [philip fernbach] at 05:47 07 June 2012

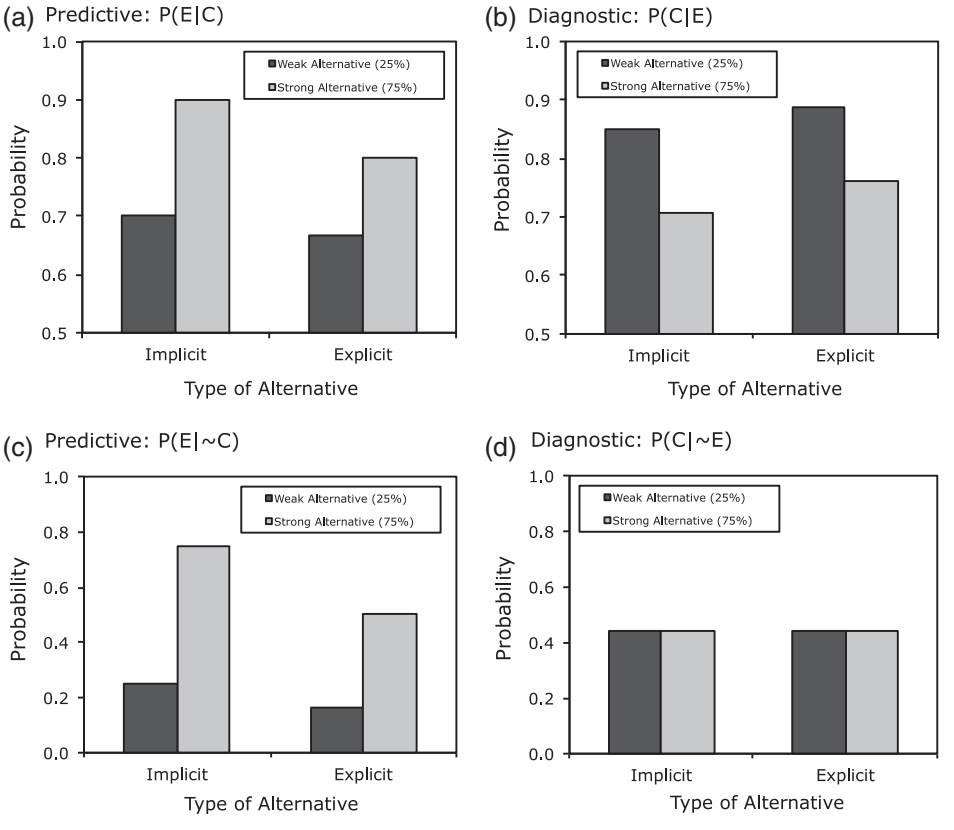

Figure 2. Normative predictions for Experiment 1. Predictions are generated assuming a base rate of 0.67 for cause features C_w and C_s. (A) Inferring an effect given the presence of its cause, (B) inferring a cause given the presence of its effect, (C) inferring an effect given the absence of its cause, and (D) inferring a cause given the absence of its effect.

Downloaded by [philip fernbach] at 05:47 07 June 2012

two artefacts (Romanian Rogos [a type of automobile] and Neptune Personal Computers). Each category had four binary feature dimensions. One value on each dimension was described as typical of the category and the other was described as atypical. For example, participants who learned Myastars were told that "Most Myastars have very hot temperature whereas some have a low temperature", "Most Myastars have high density whereas some have a low density", and so on.

Subjects were also provided with causal knowledge corresponding to the structures in Figure 1. Each causal relationship was described as one typical feature causing another, with one or two sentences describing the mechanism responsible for the causal relationship (see Table 1 for an example). In addition, a sentence describing the strength of the relationship (either 40% or 80%) was worded to convey the fact that the strength represented the power or propensity of the cause to individually produce the effect (rather than a conditional probability of the effect given the cause). For example, for the Myastar causal relationship between high density and a large number of planets, subjects were told "Whenever a Myastar has high density, it will cause that star to have a large number of planets with probability *x%*", where *x* was either 40 or 80. Note that Experiment 2 changes this wording to further emphasise the generative nature of the causal strength information.

Participants were also given information about the possibility of alternative causes of E_w and E_s. For example, participants who learned about Myastars learned not only that high density

causes a large number of planets but also that "There are also one or more other features of Myastars that cause a large number of planets. Because of this, even when its known cause (high density) is absent, a large number of planets occurs in x% of all Myastars", where x was either 25 or 75. The assignment of the four typical category features to the roles C_s, E_w, C_s, and E_s in Figure 1 was balanced over subjects, such that for each category a pair of features played the role of C_w and E_w for half the subjects and C_s and E_s for the other half. The features and causal relationships for all six categories are available from the authors.

Participants

Ninety-six New York University undergraduates received course credit for participating in this experiment. There were three between-subject factors: weak (40%) *vs.* strong (80%) focal causes, the two assignments of physical features to roles of C_w, E_w, C_s, and E_s, and which category was learned (6 levels). Participants were randomly assigned to these $2 \times 2 \times 6 = 24$ between-participant cells subject to the constraint that an equal number appeared in each cell.

Procedure

Experimental sessions were conducted by a computer. Participants first studied several screens of information that presented the category's cover story, which features occurred in "most" *vs.* "some" category members, the two causal relationships (their strength and causal mechanism), the presence of alternative causes (strengths of 25% or 75%) for features E_w and E_s, and a diagram similar to that in Figure 1. When ready, participants took a multiple-choice test that tested them on the knowledge they had just studied. While taking the test, participants were free to return to the information screens; however, doing so obligated them to retake the test.

Participants then performed inference and feature likelihood tests. During the inference test, participants were presented with two blocks of eight inference questions. Four of the eight questions involved features C_w and E_w. They were asked to predict the effect, given the presence of the cause and its absence and to predict the cause given the presence of the effect and its absence, that is, to estimate $P(E_w|C_w)$, $P(E_w|\overline{C_w})$, $P(C_w|E_w)$, and $P(C_w|\overline{E_w})$. The analogous four questions were asked for features C_s and E_s. For each question, participants were asked to suppose that a category member had been found with one feature and were asked whether it had the other feature. To attenuate memory retrieval demands, participants were provided with a printed diagram of the causal relations similar to the one of those in Figure 1 and told that "To answer these questions you should use the information about causal relationships between features of [category name] that you learned about earlier in the experiment". Responses were entered by positioning a slider on a scale where the left end was labelled "Sure that it doesn't" and the right end was labeled "Sure that it does". The position of this was scaled into the range 0–20. The presentation order of test items within a block was randomised for each participant.

During the feature likelihood rating task that followed the inference test, each of the two features on the four binary dimensions was presented on the computer screen and what proportion of all category members possessed that feature was rated by the subjects. The order of these trials was randomised for each participant. Subjects could continue to refer to the printed diagram of causal relationships during this test.

Results

Inference ratings

An initial analysis of the inference ratings revealed no effects of which of the six categories participants learned and the assignment of category features to the roles of C_w, E_w, C_s, and E_s

Downloaded by [philip fernbach] at 05:47 07 June 2012

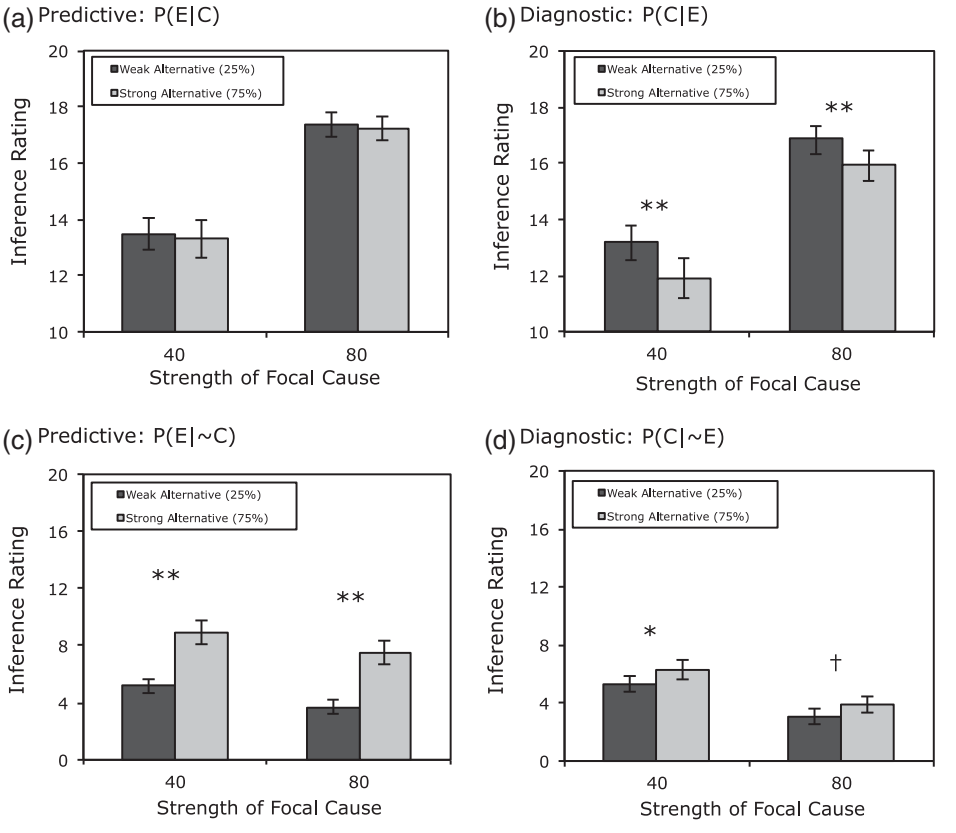

Figure 3. Inference ratings from Experiment 1 as a function of the strengths of the focal and alternative causes. (A) Inferring an effect given the presence of its cause, (B) inferring a cause given the presence of its effect, (C) inferring an effect given the absence of its cause, and (D) inferring a cause given the absence of its effect. Error bars are standard errors of the mean. $^{\dagger}p < 0.10$. $^{*}p < 0.05$. $^{**}p < 0.01$.

Downloaded by [philip fernbach] at 05:47 07 June 2012

in Figure 1. Thus, the average inference ratings collapsed over these factors and are presented in Figure 3 for each type of inference as a function of the strengths of the focal and alternative causes.

The key question of Experiment 1 concerned whether predictive inferences would be sensitive to the strength of the effect's alternative causes, and in turn whether this effect would be moderated by the strength of the focal cause. In fact, the predictive ratings shown in Figure 3(A) indicate that although subjects correctly rated the effect feature to be more likely for stronger (80%) *vs.* weaker (40%) focal causes (ratings of 17.3 *vs.* 13.4, respectively), in neither condition were the ratings at all affected by the strength of the alternative causes. A 2 × 2 mixed ANOVA with focal cause strength as the between-subject factor and alternative cause strength as the within-subject factor revealed a main effect of focal cause strength, $F(1, 94) = 31.27$, MSE $= 582$, $p < 0.0001$, confirming the larger inference ratings for the 80% focal cause, but no effect of alternative strength and no interaction, both F's <1. These results replicate those reported in Fernbach et al. (2011a) with different materials and when (a) subjects had a diagram of the causal relations (meaning those relations were highly available) and (b) when the information about alternative causes was provided in summary form.

In contrast (and also consistent with the previous results), inferences in the diagnostic direction were sensitive to the strength of the alternative causes. Figure 3(B) reveals not only that ratings were higher for stronger *vs*. weaker focal causes (average of 16.4 *vs*. 12.6 in 80% and 40% conditions, respectively), they were lower for stronger *vs*. weaker alternative causes (13.9 *vs*. 15.1), consistent with the fact that a cause is less likely when stronger alternative causes are present. A 2 × 2 ANOVA revealed a main effect of focal cause strength, $F(1, 94) = 24.67$, MSE = 712, $p < 0.0001$, a main effect of alternative strength, $F(1, 94) = 17.36$, MSE = 90, $p < 0.0001$, and no interaction, $F < 1$.

Figure 3(C) presents the predictive ratings in which the category member was explicitly stated as having the atypical value on the cause dimension (e.g. a Myastar with *low* rather than high density). These inferences were correctly sensitive to the strength of the alternative cause (average ratings of 8.2 *vs*. 4.4 in 75% and 25% conditions, respectively). A 2 × 2 ANOVA revealed a main effect of alternative cause strength, $F(1, 94) = 40.60$, MSE = 428, $p < 0.0001$. Unexpectedly, in this analysis there was a marginal effect of focal cause strength, $F(1, 94) = 3.91$, MSE = 667, $p = 0.051$, reflecting that ratings were higher for focal strengths of 80% (7.1) *vs*. 40% (5.6).[3]

Finally, Figure 3(D) presents the diagnostic ratings in which the category member was explicitly stated as having the atypical value on the effect dimension (e.g. a Myastar with a *small* number of planets). As expected, these inferences were sensitive to the strength of the focal cause (ratings of 5.8 *vs*. 3.5 in 80% and 40% conditions, respectively). A 2 × 2 ANOVA confirmed a main effect of focal cause, $F(1, 94) = 9.75$, MSE = 664, $p < 0.01$. Unexpectedly, this analysis also revealed an effect of alternative strength, $F(1, 94) = 8.58$, MSE = 114, $p < 0.01$, reflecting the fact that ratings were higher for alternative strengths of 75% (5.1) *vs*. 25% (4.2).[4]

Downloaded by [philip fernbach] at 05:47 07 June 2012

Feature likelihood ratings

The purpose of the feature likelihood test was to confirm that judgments regarding the prevalence of the effect features E_w and E_s reflected the strengths of the focal and alternative causes. The likelihood ratings for the effect features are presented in Figure 4 as a function of the two types of strengths. As expected, the effect features were rated as more prevalent both for stronger *vs*. weaker focal causes (76.6 *vs*. 67.3) and stronger *vs*. weaker alternative causes (75.8 *vs*. 68.1). A 2 × 2 ANOVA revealed a main effect of focal cause strength, $F(1, 94) = 11.62$, MSE = 353,

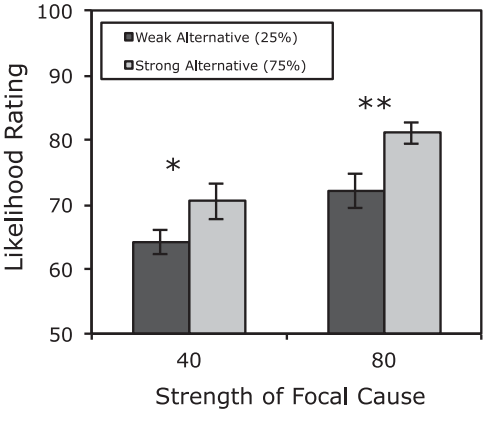

Figure 4. Likelihood ratings for the effect features (E_w and E_s in Figure 1) from Experiment 1. Error bars are standard errors of the mean. $^{*}p < 0.05$. $^{**}p < 0.01$.

$p < 0.001$, a main effect of alternative strength, $F(1, 94) = 16.80$, MSE $= 169$, $p < 0.0001$, and no interaction, $F < 1$.

Discussion

Despite the extensive differences in methodology, Experiment 1 replicated the results of Fernbach et al. (2010, 2011a) showing no sensitivity to alternative strength in predictive inferences. This pattern emerged despite the facts that providing participants with a diagram of the causal structure eliminated the need to retrieve causal relations from memory and that providing the strength of alternative causes in summary form eliminated the need to integrate over hidden variables. Diagnostic inferences, in contrast, were appropriately sensitive to the alternative causes, consistent with the asymmetry between predictive and diagnostic inferences also found in previous research.

An important question is whether the neglect of alternatives during predictive inferences reflected subjects' misunderstanding of the information we provided about alternative cause strength. Two sources of evidence argue against this possibility. First, subjects rated the effect as more likely for strong vs. weak alternative causes when the focal cause was *absent*. Second, subjects' unconditional judgments concerning the prevalence of the effect features among category members were also sensitive to the alternatives. That is, subjects were able to make use of the alternative cause information for several types of judgments, but not those involving predicting an effect given the presence of a cause.

Another concern is whether the information we provided about the strength of the focal causes was interpreted as a causal power, that is, the propensity of the cause to produce the effect. For example, neglect of alternatives would be expected if those strengths were instead interpreted as a conditional probability that incorporates the effect of alternative causes. Recall, however, that this strength information was provided as part of a description that emphasised the generative nature of the causal mechanisms and that the causal powers were written on the links on the diagrams given to participants (making it clear that they referred to the individual relations). More importantly, subjects' unconditional feature likelihood judgments were sensitive to the strength of the alternative causes showing that participants realised that the proportion of category members possessing the effect feature was higher when alternatives were strong. In Experiment 2, the causal strength information was reworded to further emphasise the generative interpretation of the strength information.

A final important result from Experiment 1 was that alternative causes were neglected during predictive inferences regardless of the strength of the focal cause. This resulted in an especially egregious judgment error in the weak focal cause condition in which an effect given the cause should be 30% more likely for a strong *vs*. weak alternative cause (Figure 2(A)). This suggests that reasoners' neglect of alternatives does not only arise when the error in judgment is likely to be small. Rather than arising from a reasoner's strategic decision to neglect alternatives, doing so appears to operate as a general heuristic that occurs regardless of the potential loss of accuracy involved.

Experiment 2

In Experiment 1 we found that reasoners fail to attend to alternative causes in predictive inferences even when the need to retrieve them from memory and integrate over their possible states was eliminated. One reason this may have occurred is that even though a representation of the alternative causes was literally right in front of them, reasoners may not have recognised their relevance to predictive inferences. That is, they may have committed what we referred to earlier as an error of commission by excising the alternative causes from the causal model with which they reasoned

Downloaded by [philip fernbach] at 05:47 07 June 2012

during predictive inferences. In Experiment 2, we assessed whether changing the representation of the alternative cause would affect participants' inferences. We varied whether the alternative was described as the net influence of "one or more other features" (the implicit condition) as in Experiment 1 or as a single explicit category feature (the explicit condition).

Participants were taught one of the category structures in Figure 5. All subjects learned categories with six binary features. Features C_w and C_s were described as causes of E_w and E_s, respectively, each with a strength of 60%. Whether or not the alternative causes of the effect features were explicit was manipulated as a between-subjects variable. In the implicit condition, the alternative causes of E_w and E_s were described as "one or more other" category features (Figure 5(A)). In the explicit condition, the alternative causes of E_w and E_s were two of the category's instructed features, namely, A_w and A_s, respectively (Figure 5(B)). Alternative cause strength was again manipulated as a within-subjects variable in both conditions, with the alternative for E_s (75%) being stronger than the alternative for E_w (25%).

Two additional changes to the materials were made. First, recall that in Experiment 1 subjects were given information regarding the likelihood of an effect when the known cause was absent (e.g. "…even when its known cause (high density) is absent, a large number of planets occurs in *x*% of all Myastars"). To make the explicit and implicit conditions of Experiment 2 comparable, analogous information was provided for the explicit alternative causes $A_w \rightarrow E_w$ and $A_s \rightarrow E_s$. For example, an alternative cause of a large number of planets in Myastars was a very hot temperature. For this causal link, subjects were not only told that "Whenever a Myastar has a very hot temperature, it will cause that star to have a large number of planets with probability 60%', but also the probability of the effect in the absence of the other cause: "This means that when a Myastar has a very hot temperature and the other cause of a large number of planets (high density) is absent, it will have a large number of planets with a probability of 60%". The focal causal relations $C_w \rightarrow E_w$ and $C_s \rightarrow E_s$ were described the same way. Note that this wording further emphasises

Downloaded by [philip fernbach] at 05:47 07 June 2012

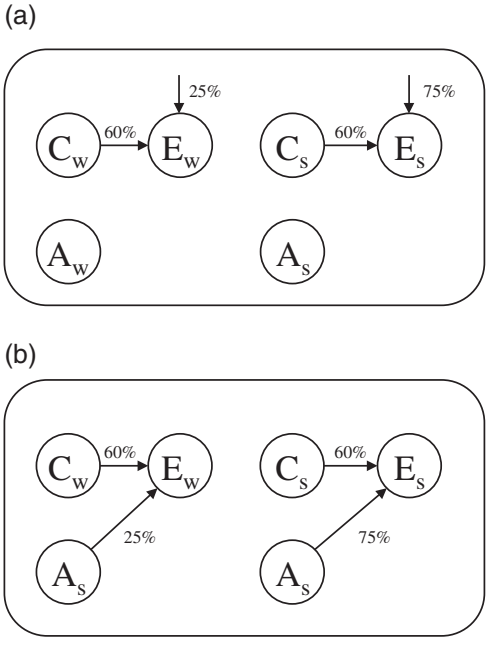

Figure 5. Causal structure tested in Experiment 2: (A) implicit alternative cause condition and (B) explicit alternative cause condition.

the generative interpretation of the causal strength information (as opposed to it being a conditional probability).

Second, recall that Experiment 1 found evidence that some subjects interpreted the causal links as having a dual sense (e.g. high density causes a large number of planets *and* low density causes a small number of planets; see Footnote 3). To address this possibility, subjects in Experiment 2 were told that most Myastars had high density and some had low density, they had either high or *normal* density. With this wording, we felt that they would be unlikely to assume a causal link between atypical feature values (e.g. that normal density causes a normal number of planets).

Although the intent of Experiment 2 is to assess whether a more explicit representation will yield greater sensitivity to alternative causes, it introduces a computational requirement that was absent in Experiment 1, namely, the need to integrate over hidden variables. Computing the influence of the alternative cause in the explicit condition requires multiplying its causal power (W_A) by the probability it is present (P_A) in the manner specified by Equation (3). Thus, making the alternative cause explicit may result in *less* sensitivity to alternative cause strength relative to the implicit condition. Experiment 3 will test conditions in which the alternative is explicit but the need for integration is avoided.

Figure 6 presents the normative predictions for Experiment 2.[5] Both predictive inferences (Figure 6(A) and (C)) and diagnostic inferences in which the effect is present (Figure 6(B)) should be sensitive to the strength of the alternative causes. However, that sensitivity should be weaker in the explicit condition (because of the need to multiply by P_{A_i}, the base rate of the alternative cause). In contrast, diagnostic inferences in which the effect is absent (Figure 6(D)) should be insensitive to the alternatives.

Downloaded by [philip fernbach] at 05:47 07 June 2012

Method

Materials

The materials were identical to those in Experiment 1 except for the two extra features and two extra causal relations required by the category structures in Figure 5, the more specific information about causal strength, and the "normal" wording for atypical features. The assignment of category features to their abstract roles (C_w, A_w, E_w, C_s, A_s, and E_s in Figure 5) was balanced over subjects so that a triple of features played the roles of C_w, A_w, and E_w for half the subjects and C_s, A_s, and E_s for the other half.

Participants

Ninety-six New York University undergraduates received course credit for participating in this experiment. There were three between-participant factors: implicit *vs*. explicit alternative causes, the two assignment of category features to the roles of C_w, A_w, E_w, C_s, A_s, and E_s, and which category was learned. Participants were randomly assigned to these $2 \times 2 \times 6 = 24$ between-participant cells subject to the constraint that an equal number appeared in each cell.

Procedure

The screens that presented the category materials and the multiple-choice test used in Experiment 1 were expanded in this experiment to include the additional two category features and (in the explicit condition) two additional causal links. Because of the larger number of questions that resulted from Experiment 2's more complicated categories, each presentation of the multiple-choice test only included those questions the subject had gotten wrong on the previous presentation.

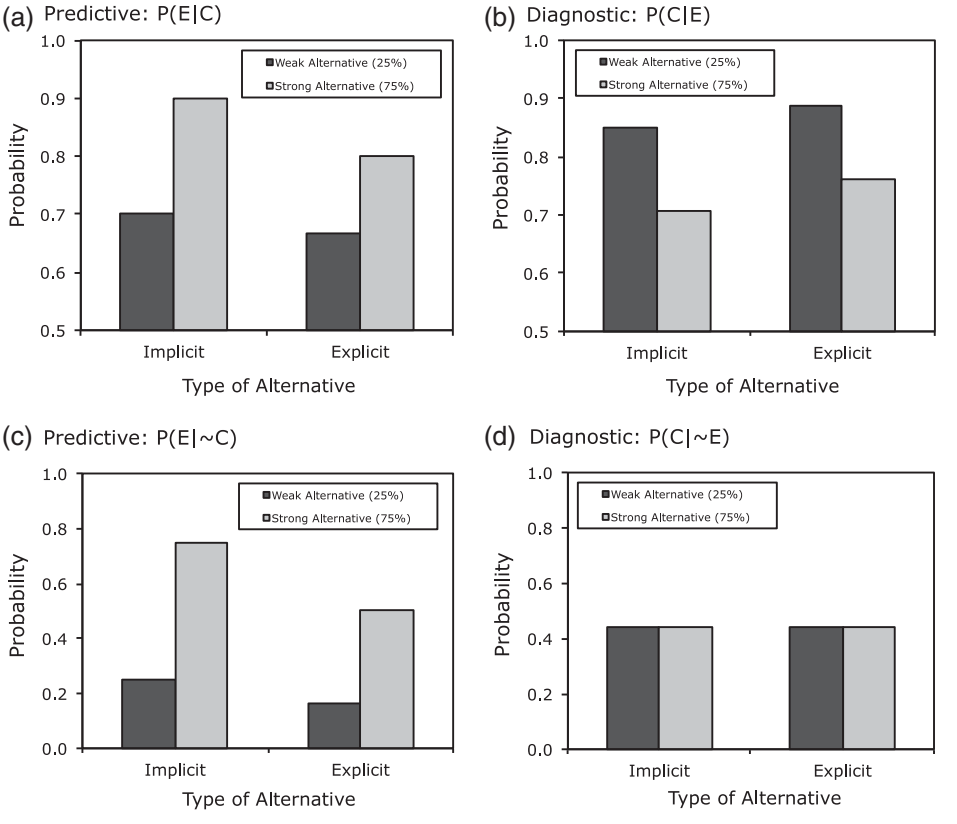

Figure 6. Normative predictions for Experiment 2. Predictions are generated assuming a base rate of 0.67 for both focal (C_w and C_s) and alternative (A_w and A_s) causes. (A) Inferring an effect given the presence of its cause, (B) inferring a cause given the presence of its effect, (C) inferring an effect given the absence of its cause, and (D) inferring a cause given the absence of its effect.

Downloaded by [philip fernbach] at 05:47 07 June 2012

As in Experiment 1, the inference test required subjects to predict the effect features given information about their causes, and to predict the causes from information about their effects. No information was provided about the state of features A_w and A_s. The feature likelihood test was identical to the one in Experiment 1 except that 12 features (2 on each of 6 binary dimensions) were presented. Subjects were again provided with a diagram of the causal links during the tests.

Results

Inference ratings

As in Experiment 1, initial analyses of the inference ratings revealed that there were no effects of which category participants learned or the assignment of category features to the roles of C_w, A_w, E_w, C_s, A_s, and E_s, and so the inference ratings are presented in Figure 7 as a function of alternative cause strength and whether the alternative causes were implicit or explicit.

The first analysis asked whether predictive inferences (Figure 7(A)) would be more sensitive to alternative cause strength when the alternative was explicit or implicit. In fact, ratings in the strong alternative condition (13.8) did not differ from those in the weak one (13.7). A 2×2 ANOVA with alternative cause type as a between-subject factor and alternative strength as a

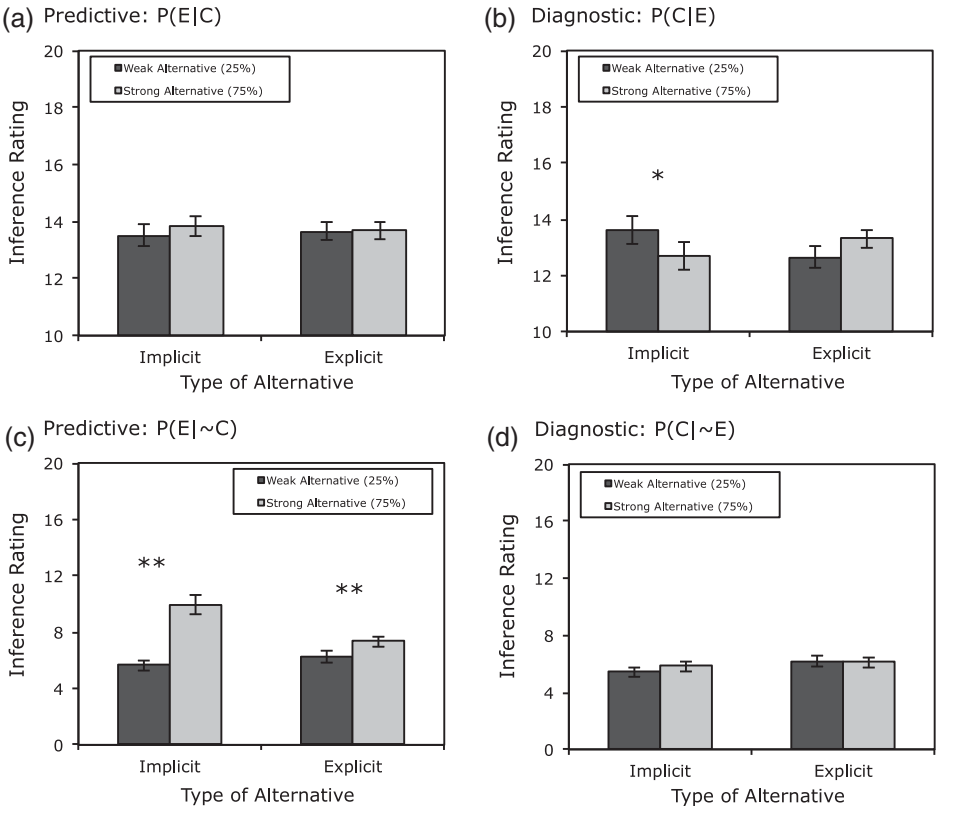

Figure 7. Inference ratings from Experiment 2 as a function of whether the alternative cause was implicit or explicit and the strength of the alternative. (A) Inferring an effect given the presence of its cause, (B) inferring a cause given the presence of its effect, (C) inferring an effect given the absence of its cause, and (D) inferring a cause given the absence of its effect. Error bars are standard errors of the mean. $^{*}p < 0.05$. $^{**}p < 0.01$.

Downloaded by [philip fernbach] at 05:47 07 June 2012

within-factor subject revealed no effect of type, $F < 1$, no effect of strength, $F(1, 94) = 1.62$, MSE $= 22$, $p > 0.20$, and no interaction, $F(1, 94) = 1.43$, MSE $= 22$, $p > 0.20$. The absence of an effect of alternative cause strength obtained in both the implicit and explicit conditions, $t(47) = 1.67$, $p = 0.17$, and $t < 1$, respectively.

Diagnostic inferences were correctly sensitive to the strength of the alternative causes in the implicit condition (Figure 7(B)). In contrast, diagnostic inferences in the explicit condition were unaffected by alternative cause strength. That is, not only did making the alternative cause an explicit category feature not result in greater sensitivity to alternative causes during predictive inferences, it made them *less* sensitive to that information during diagnostic inferences. A 2×2 ANOVA revealed no main effect alternative cause type, $F < 1$, no effect of alternative cause strength, $F < 1$, but a type $\times$ strength interaction, $F(1, 94) = 8.49$, MSE $= 86$, $p < 0.01$, reflecting the effect of alternative cause strength in the implicit but not in the explicit condition. Separate analyses found an effect of alternative cause strength in the implicit condition, $t(47) = 2.43$, $p < 0.05$, but not in the explicit condition $t(47) = 1.70$, $p = 0.09$. For the sake of brevity, the analyses of the inferences in Figure 7(C) and (D) are provided.[6]

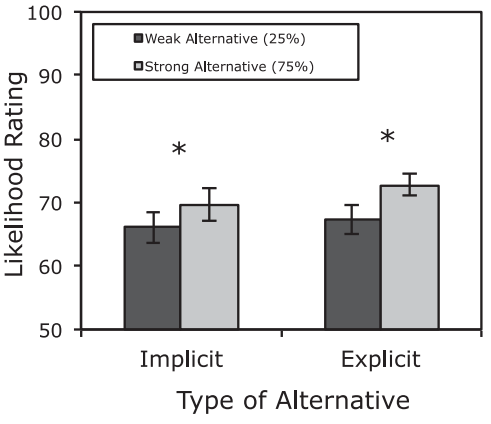

Figure 8. Likelihood ratings for the effect features (E_w and E_s in Figure 5) from Experiment 2. Error bars are standard errors of the mean. $^{*}p < 0.05$. $^{**}p < 0.01$.

Downloaded by [philip fernbach] at 05:47 07 June 2012

Feature likelihood ratings

The feature likelihood ratings confirmed that subjects' judgments of the prevalence of the effect features E_w and E_s reflected the strengths of the alternative causes. The effect feature likelihood ratings presented in Figure 8 reveal that E_s was rated as more prevalent than E_w in both the explicit and implicit conditions (71.2 *vs.* 66.7). A 2×2 ANOVA revealed a main effect of alternative strength, $F(1, 94) = 7.23, \text{MSE} = 136, p < 0.01$, but no effect alternative type and no interaction, F's <1.

Discussion

In Experiment 2, making the alternative an explicit category feature did not improve predictive judgment and made diagnostic performance worse. When predicting, participants were insensitive to alternative strength in either the explicit or implicit condition. When diagnosing, they were sensitive to alternative strength in the implicit condition only. These findings support the idea that participants failed to integrate over the possible states of the alternative cause feature. Conversely in the implicit case, the critical quantity of W_{NetAlt} – the net effect of all alternative causes – was provided in the instructions as a single parameter, making integration unnecessary.

One potential concern with this reading of Experiment 2 concerns how subjects interpreted the information we provided on the inference questions. Although we asked them to predict an effect given a cause with no information about the state of the alternative, they may have assumed that the absence of information about the alternative implied that it was known to be absent, in which case alternative strength should have no influence on the inference. Experiment 3 addresses this concern by presenting trials in which the state of the alternative cause feature is explicitly stated to be unknown.

Experiment 3

The results of Experiment 2 suggest that participants used a shortcut during integration leading to insensitivity to variation in the strength of the alternative cause. Experiment 3 further tests this possibility by again teaching subjects categories in which the alternative cause was an explicit category feature but then presenting inference questions in which the alternative cause was specified as definitely present or definitely absent. If the neglect of alternative causes in Experiment 2 was

due to the difficulty in reasoning with an alternative cause whose presence was uncertain, then subjects in Experiment 3 should show sensitivity to alternative strength on trials in which the alternative cause is known to be present.

Participants were taught the category structures in Figure 5(B). In each inference question, the state of the alternative cause variable (either A_w and A_s) was described as present or absent. For example, when subjects who learned about Myastars were asked to predict a large number of planets (an effect) given high density (a focal cause), they were also told whether the Myastar had very hot temperature (the alternative cause) or normal temperature. That is, subjects were asked to estimate $P(E_i|C_iA_i)$ and $P(E_i|C_i\overline{A_i})$. The diagnostic question was similarly asked with the alternative either present or absent: $P(C_i|E_iA_i)$ and $P(C_i|E_i\overline{A_i})$. Because this manipulation of the state of the alternative causes increased the number of questions, inferences from the absence of causes and the absence of effects were eliminated. But to ensure that the manipulation of alternative cause strength was effective, we asked subjects to make inferences involving the alternative causes themselves: $P(E_i|A_i)$ and $P(A_i|E_i)$.

Experiment 3 also addresses the alternative interpretation of Experiment 2 mentioned above, namely, that the lack of information about the state of the alternative cause feature implied that the feature was absent. It does so by also presenting trials in which the state of alternative cause feature is explicitly described as unknown. A finding that reasoners neglect alternative strength on these trials will rule out the possibility that those in Experiment 2 did so because they assumed that the alternative cause was absent.

The normative predictions for this experiment are presented in Figure 9. Figure 9(A) and (B) shows that inferences in the predictive direction should become stronger as the state of the alternative cause varies between absent, unknown, and present, and diagnostic inferences should exhibit the reverse pattern. Both predictive and diagnostic inferences should be sensitive to the strength of the alternative cause when the alternative is present or unknown (albeit that sensitivity should be lower in the unknown condition) and insensitive to that strength when the alternative is absent. Inferences from/to the alternative feature and the effect (Figure 9(C) and (D)) of course should be stronger for the stronger alternative.

Downloaded by [philip fernbach] at 05:47 07 June 2012

Method

Forty-eight New York University undergraduates received course credit for participating in this experiment. Participants were randomly assigned to the 2 (assignment of category features to roles) × 6 (category) = 12 between-participant cells subject to the constraint that an equal number appeared in each cell. The materials and procedure were identical to those used in the explicit condition of Experiment 2 except for the different questions asked during the inference test. Subjects were provided with a diagram of the causal links during both the inference and feature likelihood tests.

Results

Inference ratings

Analyses of the inference ratings again revealed that there were no effects of which category participants learned or the assignment of category features to their abstract roles. The inference ratings collapsed over these factors are presented in Figure 10.

Our central question was whether causal inferences would be affected by alternative features whose state was explicitly known to be present or absent. Examining predictive inferences first, ratings were much higher when the alternative was present as compared to when absent, 17.8 *vs.* 12.4, respectively. These results show that reasoners readily attend to alternative causes when the

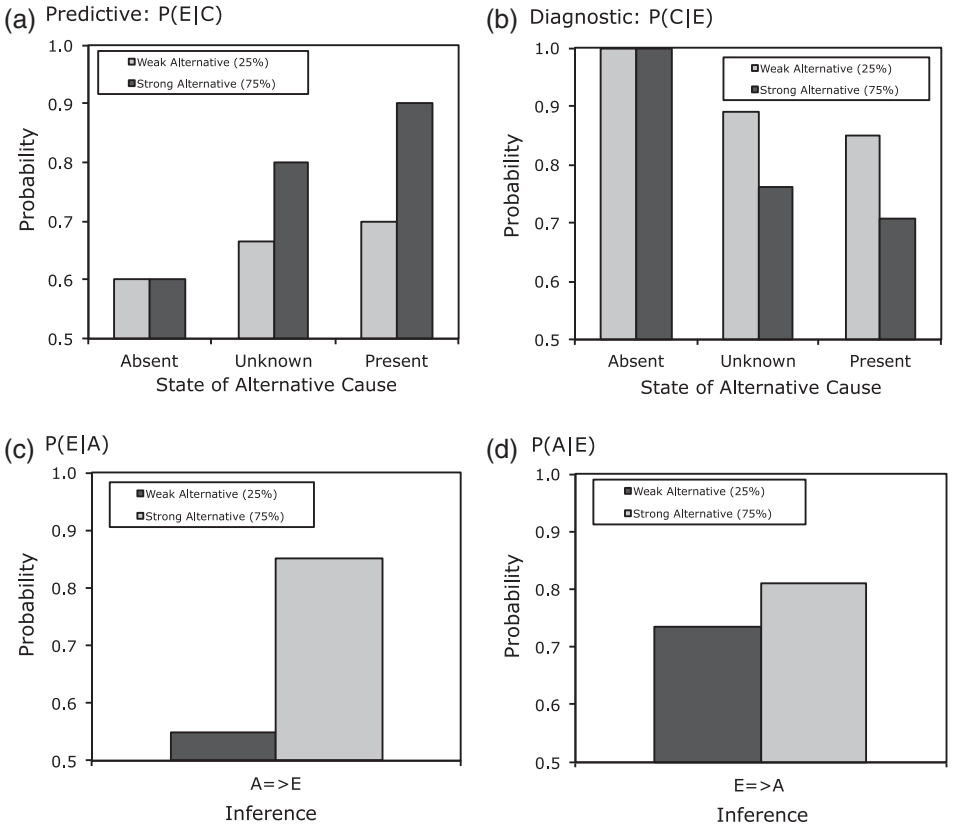

Figure 9. Normative predictions for Experiment 3. Predictions are generated assuming a base rate of 0.67 for both focal (C_w and C_s) and alternative (A_w and A_s) causes. (A) Inferring an effect given the presence of its cause, (B) inferring a cause given the presence of its effect, (C) inferring the effect feature from the alternative and (D) inferring the alternative feature from the effect.

Downloaded by [philip fernbach] at 05:47 07 June 2012

state of those causes is known. More importantly, inference ratings were significantly higher for strong *vs*. weak alternatives when the alternative was present (18.3 *vs*. 17.2). This is the only experimental condition tested in this article in which predictive inferences were sensitive to the strength of alternative causes. When the alternative was described as absent, subjects correctly showed no sensitivity to alternative strength. A 3 × 2 ANOVA of the data in Figure 10(A) with alternative cause state and strength as within-subject factors revealed a main effect of state, $F(2, 94) = 165.24$, MSE = 129, $p < 0.0001$, a marginal effect of strength, $F(1, 47) = 2.61$, MSE = 77, $p = 0.11$, but a state by strength interaction $F(2, 94) = 4.24$, MSE = 72, $p < 0.05$. Regarding the main effect of state, *t*-tests revealed that inference ratings were higher when the alternative feature was present *vs*. unknown, $t(47) = 13.01$, $p < 0.0001$, which in turn were only marginally higher than when the alternative was absent, $t(47) = 1.86$, $p = 0.07$. Regarding the state × strength interaction, there was an effect of alternative strength when the alternative was present, $t(47) = 3.17$, $p < 0.01$, but not when it was unknown, $t(47) = 1.10$, or absent, $t(47) = 1.18$, both *p*'s >0.20.

Diagnostic inferences shown in Figure 10(B) revealed an analogous pattern. Ratings were higher when the alternative was absent as compared to when present, 14.9 *vs.* 12.8, indicating that reasoners were also sensitive to the state of the alternative in diagnostic inferences. Moreover, when the state of the alternative feature was present, these inferences were appropriately

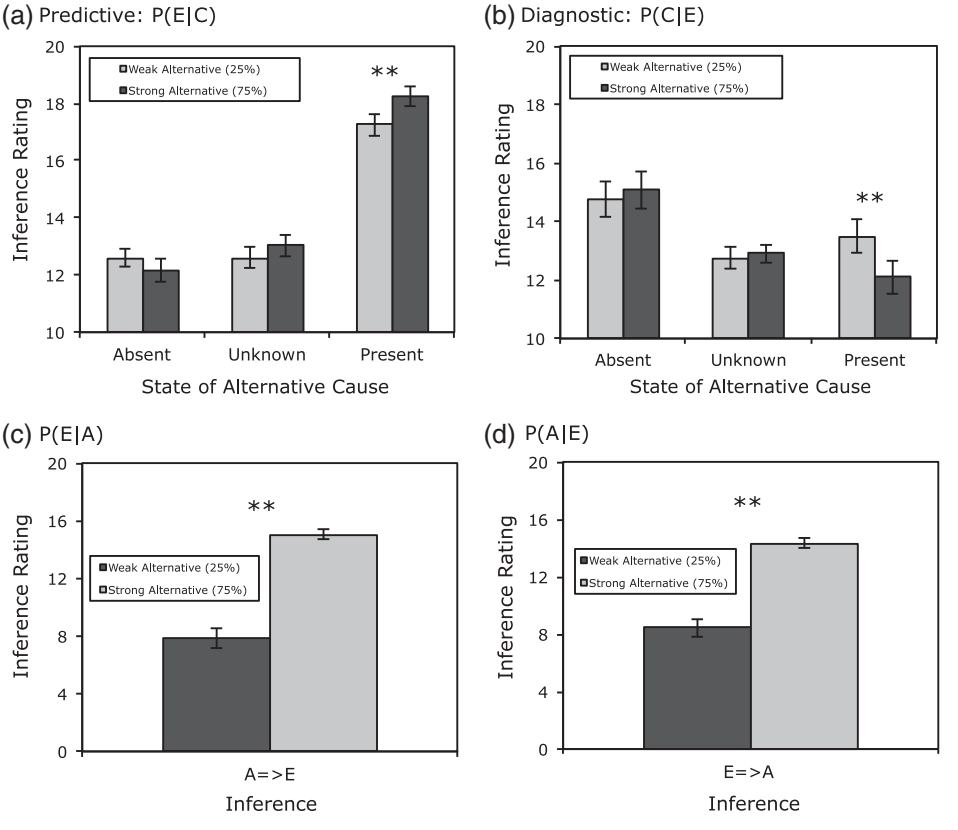

Figure 10. absent, present, or unknown. (A) Inferring an effect given the presence of its cause. (B) Inferring a cause given the presence of its effect. (C) Inferring the effect feature from the alternative. (D) Inferring the alternative feature from the effect. Error bars are standard errors of the mean. $^{\dagger}p < 0.10$. $^{*}p < 0.05$. $^{**}p < 0.01$.

Downloaded by [philip fernbach] at 05:47 07 June 2012

stronger for weak *vs*. strong alternatives (13.5 *vs.* 12.1). Subjects correctly showed no sensitivity to alternative strength when the alternative feature was absent. A 3×2 ANOVA of the data in Figure 10(B) found a main effect of alternative state, $F(2, 94) = 8.06, \text{MSE} = 437, p < 0.001$, no effect of alternative strength, $F(1, 47) = 1.51, \text{MSE} = 122, p > 0.20$, but a significant interaction, $F(2, 94) = 4.13, \text{MSE} = 132, p < 0.01$. Regarding the main effect, inference ratings were higher when the alternative was absent *vs*. unknown, $t(47) = 4.18$, $p < 0.0001$, which in turn did not differ from when the alternative was present, $t < 1$. Regarding the interaction, there was an effect of alternative strength when the alternative was present, $t(47) = 3.33$, $p < 0.01$, but not when it was unknown or absent, both t's <1.

Experiment 3 also tested whether inferences would be sensitive to alternative strength when the state of the alternative cause was explicitly stated to be unknown. In fact, replicating Experiment 2, neither predictive (Figure 10(A)) nor diagnostic (Figure 10(B)) inference ratings showed sensitivity to alternative strength in this condition. Instead, ratings for these problems were very similar to those in Experiment 2's explicit condition in which no information was provided about the state of the alternative.

Finally, the inferences involving the alternative cause feature itself showed the expected pattern. Ratings were higher for the stronger *vs*. weaker alternative causal link both when effect was inferred

from the alternative (15.1 *vs.* 7.8, Figure 10(C)), $t(47) = 9.31$, $p < 0.0001$, and vice versa (14.4 *vs*. 8.5, Figure 10(D)), $t(47) = 8.24$, $p < 0.0001$.

Feature likelihood ratings

The feature likelihood ratings again confirmed that subjects' judgments regarding the prevalence of the effect features reflected the strengths of the alternative causes. The effect feature with the stronger alternative cause was rated to be more prevalent than the one with the weaker alternative (73.9 *vs*. 69.4), $t(47) = 2.07$, $p < 0.05$.

Discussion

Experiment 3 established two important findings: First, the sensitivity to the strength of an alternative cause during diagnostic inferences, lost in Experiment 2, was restored so long as the alternative was definitively known to be present. Second, predictive inferences were also sensitive to alternative strength when the alternative was present. This result shows that people are not intrinsically unable to take alternatives into account when making predictions. Rather, it suggests that neglecting alternatives is due to errors in representational and computational steps leading up to the final judgment.

Downloaded by [philip fernbach] at 05:47 07 June 2012

General discussion

Summary of results

Over three experiments we examined the types of errors that people make when judging predictive and diagnostic probability. We summarise the results by spelling out three conclusions that emerge from the current studies. First, the neglect of alternative causes is not due solely to the difficulty in retrieving those causes from memory. Evidence for this conclusion comes from all three experiments. Participants were instructed on the causal links as part of the experimental session and then provided a diagram of those relations during the inference task, but they still made errors. This does not mean that errors in causal reasoning would not be even worse in cases where alternative causes are hard to retrieve; we suspect they would be. But our results do show that although availability in memory of alternative causes may be a necessary condition for veridical causal reasoning, it is by no means a sufficient one.

Second, neglecting alternatives is also not due to strategic laziness, that is, relaxing reasoning norms only when the potential judgment error is small. Evidence for this conclusion comes from Experiment 1, which manipulated the strength of the focal cause. Although we showed that the magnitude of the error in predictive inferences, due to ignoring alternative causes, increases as the strength of the focal cause decreases, participants in Experiment 1 ignored alternative causes even when the focal cause was weak, resulting in especially egregious judgment errors. The violations of the normative model observed in Experiments 2 and 3 were also substantial in magnitude. In this regard at least, causal reasoning errors seem to share the hallmarks of heuristics, shortcuts that are applied liberally and unconsciously – and sometimes lead to serious errors.

Third, alternative causes are more likely to be ignored when there is uncertainty about either the state or identity of those causes. Our evidence for the importance of the *state* of an alternative cause comes from a comparison of trials in Experiment 3 in which the state of the alternative was either present or unknown: Reasoners were sensitive to the strength of the alternative causes for the former but not the latter. We believe that knowledge of the state of the alternative results in better reasoning because the need to multiply the strength of an alternative (W_A in Equation (3)) by the probability that the cause is present (P_A) is eliminated. Our evidence for the importance of the *identity* of the alternative causes comes from Experiments 1 and 2 that showed that alternatives

were neglected even when their influence was provided in a summary form (i.e. as the net effect of "one or more" other causes), thus eliminating the need to compute $P_A W_A$. We believe that knowledge of the identity of the alternative results in better reasoning because reasoners are less likely to commit an error of commission by excising the alternatives from the model.

Taken together, these results support the view that people find it easier to make causal inferences when relevant knowledge can be represented simply and concretely. Situations in which variables or their states are unidentified invite reasoners to reduce effort by invoking shortcuts that can lead to error. These findings are consistent with other research showing that reasoners attend to alternative causes when all relevant causal factors are known with certainty (Rehder and Burnett 2005). It also dovetails with developmental experiments by Fernbach et al. (2011) showing that 3-year-olds are capable of diagnostic inference with known causes, but that diagnostic reasoning with uncertain causes only emerges at age 4.

Directions for future research

We have shown how specific difficulties in the representational and computational requirements of causal inference can lead to errors. But rather than abandon the normative model, we advocate an approach that identifies those subcomponents of causal reasoning that are representationally and computationally demanding and thus invite reasoners to take shortcuts. This approach has been fruitfully applied to explicate failures to adhere to the normative rules of causal learning (De Houwer and Beckers 2003; Reips and Waldmann 2008; Waldmann and Walker 2008). A promising avenue for future research would be to make use of the sort of processing load manipulations used in these learning studies to assess whether they also have the expected effects on causal inferences. Another would be to strive for an integrated account of how hidden alternative causes influence both reasoning and learning (cf. Hagmayer and Waldmann 2007; Luhmann and Ahn 2007).

Besides those we have investigated, there are of course numerous other task variables that might influence the veridicality of people's causal inferences. For example, to assess reasoning with causal models that are fully specified, we provided causal strength information in the form of explicit probabilities. Yet, the difficulty that people have reasoning with probabilities (as compared to natural frequencies) is well documented (Gigerenzer and Hoffrage 1995; also see Barbey and Sloman 2007). Thus, one might ask whether reasoning would improve if causal strength information were presented in a format that people reason with more naturally.[7]

Finally, although people did not exhibit strategic laziness (i.e. only taking shortcuts that lead to minimal error) in this study, this does not rule out the possibility that they have learned from experience which types of shortcuts usually result in little loss of accuracy. Indeed, the basic asymmetry between predictive and diagnostic inferences (originally documented by Fernbach et al. 2010; 2011a, and replicated here) is explicable in terms of the potential errors usually associated with these classes of inference problems. On the one hand, the magnitude of potential error due to neglecting alternatives (i.e. incorrectly treating W_{NetAlt} in Equation (1) as if it's 0) during predictive inferences will often be small because it is bounded between the power of the focal cause and one. In contrast, if alternatives are ignored during diagnostic reasoning then the focal cause *must* be present (Equation (2)). Because considering even a weak or improbable alternative can thus sharply change a diagnostic inference, people may have learned from experience that ignoring alternative causes in diagnostic inferences usually leads to more serious errors than predictive ones.

Conclusion

We conclude by emphasising the empirical successes of causal Bayes nets in many cognitive domains, including inference, analogy, classification, generalisation, and learning. These results

Downloaded by [philip fernbach] at 05:47 07 June 2012

provide compelling evidence that human thinking goes beyond the associative- and similarity-based process that dominate many theories, past and present. But as sophisticated as such thinking might be, it is still vulnerable to the human desire to reduce mental effort. This research represents an attempt to understand reasoning errors *within* the normative causal Bayes net framework. It is true that Bayes nets cannot explain errors of reasoning, but we do believe this framework is a promising avenue for research because people are generally good causal reasoners. An apt analogy would be to a talented builder with good tools and a few bad habits. Focusing on the bad habits to the exclusion of everything else leads to a false impression. But that does not mean we should let the bad habits slide.

Downloaded by [philip fernbach] at 05:47 07 June 2012

Notes

1. Probabilities shown in Figure 2 were generated assuming the base rate of the cause features (P_c) within category members was 0.67. Thus, because subjects were not given any information about feature base rates, the predictions are ordinal only (they only reflect the relative strength of each kind of inference).
2. The probability of the cause given the absence of the effect is given by
$$P(\text{Cause}|\overline{\text{Effect}}) = \frac{P_C - P_C W_C}{1 - P_C W_C}.$$
3. This effect is nonnormative if one interprets the causal links as having a *single sense*, that is, if (in Myastars for example) high density causes a large number of planets but not that *low* density causes a *small* number of planets. Under this interpretation, focal cause strength should not affect $P(E|\overline{C})$ inferences (e.g. the strength of the causal link between high density and a large number of planets is irrelevant for those Myastars with low density). However, this effect can be understood if a minority of subjects interpreted the causal links as having a *dual sense*, that is, if "high density causes a large number of planets" also meant that "low density causes a small number of planets". In this case, the focal cause strength becomes relevant to $P(E|\overline{C})$ inferences (E is less likely as strength increases because not-C is more likely to produce not-E). We will address this possible dual sense interpretation of the causal links with a minor change in the wording of the materials in Experiment 2.
4. This effect is not predicted by the normative model (under either the dual or single sense interpretations of the causal links) or by the effort reduction framework. It did not replicate in the following experiment and so it will not be discussed further.
5. As in Experiment 1, these (ordinal) predictions were generated assuming that $P_c = 0.67$. In addition, predictions in the explicit condition were generated assuming a base rate of 0.67 for each alternative cause and that each E_i had no causes other than C_i and A_i.
6. Figure 7(C) shows that $P(E|\overline{C})$ inferences were appropriately sensitive to the strength of alternative causes in both the implicit and explicit conditions, although this effect was larger in the implicit condition (9.9 *vs.* 5.6) than the explicit one (7.4 *vs.* 6.3), consistent with the normative model (Figure 6(C)). A 2×2 ANOVA revealed an effect of alternative strength, $F(1, 94) = 44.61$, MSE $= 194$, $p < 0.0001$, a marginal effect of alternative cause type, $F(1, 94) = 3.37$, MSE $= 357$, $p = 0.07$, and a type $\times$ strength interaction, $F(1, 94) = 15.89$, MSE $= 194$, $p < 0.0001$, reflecting the larger effect of alternative cause strength in the implicit condition. Finally, Figure 7(D) shows that inferences were correctly insensitive to alternative cause strength; an ANOVA revealed no main effects and no interaction, all p's >0.20.
7. We thank Michael Waldmann for suggesting this manipulation.

References

Acevedo, M., and Krueger, J.I. (2005), 'Evidential Reasoning in the Prisoner's Dilemma', *American Journal of Psychology*, 18, 431–457.

Barbey, A.K., and Sloman, S.A. (2007), 'Base-rate Respect: From Ecological Rationality to Dual Processes', *Behavioral and Brain Sciences,* 30, 241–297.

Beckers, T., De Houwer, J., Miller, R.R., and Urushihara, K. (2006), 'Reasoning Rats: Forward Blocking in Pavlovian Animal Conditioning is Sensitive to Constraints of Causal Inference', *Journal of Experimental Psychology: General,* 135, 92–102.

Blaisdell, A.P., Sawa, K., Leising, K.J., and Waldmann, M.R. (2006), 'Causal Reasoning in Rats', *Science,* 311, 1020–1022.

Cheng, P. (1997), 'From Covariation to Causation: A Causal Power Theory', *Psychological Review,* 104, 367–405.

De Houwer, J., and Beckers, T. (2003), 'Secondary Task Difficulty Modulates Forward Blocking in Human Contingency Learning', *The Quarterly Journal Of Experimental Psychology,* 56B, 345–357.

Fernbach, P.M., Darlow, A., and Sloman, S.A. (2010), 'Neglect of Alternative Causes in Predictive But not Diagnostic Reasoning', *Psychological Science,* 21, 329–336.

Fernbach, P.M., Darlow, A., and Sloman, S.A. (2011a), 'Asymmetries in Predictive and Diagnostic Reasoning', *Journal of Experimental Psychology: General,* 140, 168–185.

Fernbach, P.M., Darlow, A., and Sloman, S.A. (2011b), 'When Good Evidence Goes Bad: The Weak Evidence Effect in Judgment and Decision-Making. *Cognition,* 119, 459–467.

Fernbach, P. M., Macris, D. M. and Sobel, D. M. (2012). 'Which one made it go? The emergence of diagnostic reasoning in preschoolers', *Cognitive Development*, 27 (1), 39–53.

Fernbach, P.M., and Sloman, S.A. (2009), 'Causal Learning with Local Computations', *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 35(3), 678–693.

Fernbach, P.M, Sobel, D.A., and Macris, D. (2011), 'Which One Made It Go? The Emergence of Diagnostic Reasoning in Preschoolers', *Cognitive Development,* in press.

Fischoff, B. and Beyth, R. (1975), "'I Knew It Would Happen": Remembered Probabilities of Once Future Things', *Organizational Behavior and Human Performance*, *13*, 1–16.

Gigerenzer, G., and Hoffrage, U. (1995), 'How to Improve Bayesian Reasoning Without Instruction: Frequency Formats', *Psychological Review,* 102, 684–704.

Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., and Kushnir, T. (2004), 'A Theory of Causal Learning in Children: Causal Maps and Bayes Nets', *Psychological Review,* 111, 3–23.

Glymour, C. (1998), 'Learning Causes: Psychological Explanations of Causal Explanation', *Minds and Machines*, 8, 39–60.

Griffiths, T.L., and Tenebaum, J.B. (2005), 'Structure and Strength in Causal Induction', *Cognitive Psychology,* 51, 334–384.

Griffiths, T.L., and Tenenbaum, J.B. (2009), 'Theory-Based Causal Induction', *Psychological Review,* 116, 661–716.

Hagmayer, Y., and Waldmann, M.R. (2007), 'Inferences About Unobserved Causes in Human Contingency Learning', *The Quarterly Journal of Experimental Pscyhology,* 60, 330–355.

Hayes, B.K., and Thompson, S.P. (2007), 'Causal Relations and Feature Similarity in Children's Inductive Reasoning', *Journal of Experimental Psychology: General,* 107, 470–484.

Hayes, B.K., and Rehder, B. (in press). 'Causal Categorization in Children and Adults', *Cognitive Science*.

Holyoak, K.J. and Cheng, P.W. (2011), 'Causal Learning and Inference as a Rational Process', *Annual Review of Psychology*, *62*, 135–163.

Holyoak, K.J., Lee, J.S., and Lu, H. (2010), 'Analogical and Category-Based Inferences: A Theoretical Integration With Bayesian Causal Models', *Journal of Experimental Psychology: General,* 139, 702–727.

Jordan, H.I. (Ed.). (1999), *Learning in Graphical Models*. Cambridge, MA: MIT Press.

Kahneman, D. and Frederick, S. (2002), 'Representative Revisited: Attribute Substitution in Intuitive Judgment', *Heuristics and Biases: The psychology of Intuitive Judgment*, eds. T. Gilovich, D. Griffin and D. Kahneman, New York: Cambridge University Press, pp. 49–81.

Kahneman, D., and Tversky, A. (1982), 'The Simulation Heuristic', (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*, eds. D. Kahneman, P. Slovic and A. Tversky, New York, NY: Cambridge University Press, pp. 201–208.

Kemp, C., and Tenenbaum, J.B. (2009), "Structured Statistical Models of Inductive Reasoning', *Psychological Review,* 116, 20–58.

Langer, E.J. (1975), 'The Illusion of Control', *Journal of Personality and Social Psychology*, 32, 311–328.

Downloaded by [philip fernbach] at 05:47 07 June 2012

Lee, H.S., and Holyoak, K.J. (2008), 'The Role of Causal Models in Analogical Inference', *Journal of Experimental Psychology: Learning, Memory and Cognition,* 34, 1111–1122.

Lu, H., Yuille, A.L., Liljeholm, M., Cheng, P.W., and Holyoak, K.J. (2008), Bayesian generic priors for causal learning. *Psychological Review,* 115, 955–984.

Luhmann, C.C., and Ahn, W. (2007), 'BUCKLE: A Model of Unobserved Cause Learning', *Psychological Review,* 114, 657–677.

Novick, L.R., and Cheng, P.W. (2004), 'Assessing Interactive Causal Influence', *Psychological Review,* 111, 455–485.

Opfer, J.E., and Bulloch, M.J. (2007), 'Causal Relations Drive Young Children's Induction, Naming, and Categorization', *Cognition,* 105, 206–217.

Payne, J.W., Bettman, J.R., and Johnson, E.J. (1993), *The Adaptive Decision Maker*, New York: Cambridge University Press.

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, Cambridge, UK: Cambridge University Press.

Quattrone, G.A., and Tversky, A. (1984), 'Causal Versus Diagnostic Contingencies: On Self-deception and the Voter's Illusion', *Journal of Personality and Social Psychology*, 46, 237–248.

Rehder, B. (2003a), 'Categorization As Causal Reasoning', *Cognitive Science,* 27, 709–748.

Rehder, B. (2003b), 'A Causal-Model Theory of Conceptual Representation and Categorization', *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 29, 1141–1159

Rehder, B. (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., and Burnett, R.C. (2005), 'Feature Inference and the Causal Structure of Object Categories', *Cognitive Psychology,* 50, 264–314.

Rehder, B., and 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., and Hastie, R. (2004), 'Category Coherence and Category-Based Property Induction', *Cognition,* 21, 113–153.

Rehder, B., and Kim, S. (2006), 'How Causal Knowledge Affects Classification: A Generative Theory of Categorization', *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 32, 659–683.

Rehder, B., and Kim, S. (2009), 'Classification as Diagnostic Reasoning', *Memory & Cognition,* 37, 715–729.

Rehder, B., and Kim, S. (2010), 'Causal Status and Coherence in Causal-Based Categorization', *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1171–1206.

Reips, U., and Waldmann, M.R. (2008), 'Sensitivity to Base Rates: Challenges for Theories of Causal Learning', *Experimental Psychology,* 55, 9–22.

Risen, J.L., and Gilovich, T. (2008), 'Why People Are Reluctant to Tempt Fate', *Journal of Personality and Social Psychology*, 95, 293–307.

Rottenstreich, Y., and Tversky, A. (1997), 'Unpacking, Repacking and Anchoring: Advances in Support Theory', *Psychological Review*, 104, 406–415.

Shafto, P., Kemp, C., Bonawitz, E.B., Coley, J.D., and Tenebaum, J.B. (2008), 'Inductive Reasoning About Causally Transmitted Properties', *Cognition,* 109, 175–192.

Shah, A.K., and Oppenheimer, D.M. (2008), 'Heuristics Made Easy: An Effort Reduction Framework', *Psychological Bulletin, 134*, 207–222.

Sloman, S.A. (2005), '*Causal Models; How People Think About the World and Its Alternatives*', New York: Oxford University Press.

Sloman, S.A., Fernbach, P.M., and Hagmayer, Y. (2010), 'Self Deception Requires Vagueness', *Cognition,* 115, 268–281.

Sobel, D.M., Tenenbaum, J.B., and Gopnik, A. (2004), 'Children's Causal Inferences From Indirect Evidence: Backwards Blocking and Bayesian Reasoning in Preschoolers', *Cognitive Science*, 28, 303–333.

Spirtes, P., Glymour, C., and Scheines, R. (1993), *Causation, Prediction, and Search*, New York: Springer-Verlag.

Swirsky, C., Fernbach, P.M., and Sloman, S.A. (2011), 'An Illusion of Control Modulates the Reluctance to Tempt Fate', *Judgment and Decision Making*, 6, 688–696.

Downloaded by [philip fernbach] at 05:47 07 June 2012

Tversky, A., and Kahneman, D. (1980), 'Causal Schemata in Judgments Under Uncertainty', *Progress in Social Psychology*, ed. M. Fishbein, Hillsdale, NJ: Erlbaum, pp. 49–72.

Tversky, A., and Kahneman, D. (1983), 'Extensional vs. Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement', *Psychological Review,* 91, 293–315.

Tversky, A., and Koehler, D.J. (1994), 'Support Theory: A Nonextensional Representation of Subjective Probability', *Psychological Review*, 101, 547–567.

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., Cheng, P.W., Hagmayer, Y., and Blaisdell, A.P. (2008), 'Causal Learning in Rats and Humans: A minimal Rational Model', in *The Probabilistic Mind. Prospects for Bayesian Cognitive Science,* eds. N. Chater and M. Oaksford, Oxford: Oxford: University Press, pp. 453–484.

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

Waldmann, M.R., and Walker, J.M. (2005), 'Competence and Performance in Causal Reasoning', *Learning & Behavior,* 33, 211–229.

Wells, G.L., and Gavanski, I. (1989), 'Mental Simulation of Causality', *Journal of Personality and Social Psychology*, 56, 161–169.

Downloaded by [philip fernbach] at 05:47 07 June 2012