Journal of Experimental Psychology: General

0096-3445/10/$12.00 DOI: 10.1037/a0022100

Asymmetries in Predictive and Diagnostic Reasoning

Philip M. Fernbach, Adam Darlow, and Steven A. Sloman
Brown University

In this article, we address the apparent discrepancy between causal Bayes net theories of cognition, which posit that judgments of uncertainty are generated from causal beliefs in a way that respects the norms of probability, and evidence that probability judgments based on causal beliefs are systematically in error. One purported source of bias is the ease of reasoning forward from cause to effect (*predictive reasoning*) versus backward from effect to cause (*diagnostic reasoning*). Using causal Bayes nets, we developed a normative formulation of how predictive and diagnostic probability judgments should vary with the strength of alternative causes, causal power, and prior probability. This model was tested through two experiments that elicited predictive and diagnostic judgments as well as judgments of the causal parameters for a variety of scenarios that were designed to differ in strength of alternatives. Model predictions fit the diagnostic judgments closely, but predictive judgments displayed systematic neglect of alternative causes, yielding a relatively poor fit. Three additional experiments provided more evidence of the neglect of alternative causes in predictive reasoning and ruled out pragmatic explanations. We conclude that people use causal structure to generate probability judgments in a sophisticated but not entirely veridical way.

Keywords: judgment, causal reasoning, prediction, diagnostic reasoning, causal models

A consensus is emerging that most, if not all, of our beliefs are probabilistic in the sense that they come in degrees (Chater & Oaksford, 2008). Many of these uncertain beliefs are grounded in causal knowledge, an understanding of how causes lead to effects (Gopnik & Schulz, 2007). But what precisely is the relation between our causal beliefs and judgments of probability? A growing literature in both philosophy and psychology argues that probability judgments are generated from causal beliefs that accord with probabilistic norms. However, evidence from the psychology of judgment suggests that probability judgments based on causal structure are systematically in error, implying that such normative theories are insufficient to account for human judgment. Our goal of this article was to adjudicate between these possibilities by evaluating the extent to which beliefs about causal structure give rise to judgments of probability.

Causal Bayes Nets

A recent advance in relating causal knowledge with probability judgment is the development of causal Bayes nets, a normative framework in which a causal structure is used to define a probability distribution (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993). Causal Bayes nets are graphical representations of probability distributions with two components: (a) a graph composed of nodes and arrows that represent relations of probabilistic dependence among the variables of a causal system such that arrows point from causes to effects and (b). a set of conditional probabilities on each node that represent the likelihood of each effect, given all possible patterns of causes. Causal Bayes net theories of cognition posit that judgments of probability arise from a representation that approximates the structure of the causal system generating the property or event being judged. To illustrate, individuals judge the probability of arriving home on time by taking into account the chain of causes that comprise the path home (e.g., walking to the car, driving to the highway), possible disablers (e.g., the car will not start), and necessary enablers (e.g., that the car has sufficient fuel). These variables sit in a highly structured relation. A causal model is a representation of that structure that affords the ability to compute a coherent probability.

Several recent psychological theories have suggested that causal Bayes nets serve as the fundamental representational building block for cognition (Glymour, 2001; Gopnik et al., 2004; Sloman, 2005). Causal Bayes nets have been the basis for theories of category knowledge and induction (Rehder, 2003), learning (Anderson, 1990; Griffiths & Tenenbaum, 2005; Sobel, Tenenbaum, & Gopnik, 2004; Waldmann & Holyoak, 1992), decision making (Sloman & Hagmayer, 2006), conditional inference (Sloman & Lagnado, 2005), the meaning of causal words (Sloman, Barbey, & Hotaling, 2009), and intentionality judgment (Sloman, Fernbach, & Ewing, 2010). Each of these theories comes equipped with supporting data. Further supporting data comes from the fact that some classic fallacies of probability judgment, like base-rate neglect, appear to diminish when task instructions make the causal structure underlying the inference

Philip M. Fernbach, Adam Darlow, and Steven A. Sloman, Department of Cognitive, Linguistic, and Psychological Sciences, Brown University.

This work was supported by a Galner Dissertation Fellowship and an American Psychological Association Dissertation Research Award to Philip M. Fernbach and by National Science Foundation Award 0518147 to Steven A. Sloman. Experiment 1 and aspects of the model were published in the *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, which was held in July 2009 in Austin, TX. We thank Jonathan Bogard for help in collecting data and David Over, Dinos Hadjichristidis, and Pat Shafto for helpful discussions of the work.

Correspondence concerning this article should be addressed to Philip M. Fernbach, Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Box 1821, Providence, RI 02912. E-mail: philip_fernbach@brown.edu

clearer (Ajzen, 1977; Krynski & Tenenbaum, 2007; Tversky & Kahneman, 1982).

Nonnormative Causal Inference

Despite these developments, longstanding phenomena in the probability judgment literature pose a problem for causal Bayes nets as a descriptive theory. People show systematic biases and neglect relevant information when making probability judgments based on causal information. These phenomena suggest that the way people reason with causal information is not normative and therefore not consistent with causal Bayes nets. The most direct evidence comes from seminal work by Tversky and Kahneman (1980). They asked people to compare the conditional probabilities of events while varying the type of causal relation, and they attempted to hold constant the evidential relation between the evidence and the event to be judged so that any differences must be due to the influence of the type of causal relation on judgment. Their key manipulation was based on the fundamentally asymmetric nature of causality: causes generate their effects, but not vice versa. In contrast, merely evidential relations hold in both directions. In criminal trials, for example, evidence may consist of arguments about motive and physical evidence. Motive refers to properties or events that may have caused the defendant to commit the crime. This is called *predictive evidence*, because it could have been used to predict that the crime would occur. Physical evidence is a property or event (e.g., fingerprints found on a weapon) that is a possible effect of the commission of the crime. This is called *diagnostic evidence*.

Tversky and Kahneman (1980) hypothesized that because it is easier and more natural to think in the direction of causality than against it, people would judge predictive inferences greater than diagnostic inferences, all else being equal, and would also be more confident in them. Their idea was that probability "flows" mentally from cause to effect, the way that water flows down an incline. Their results provided supportive evidence. For instance, people judged the probability that a daughter has blue eyes given that her mother does (a predictive inference) to be greater than the probability that a mother has blue eyes given her daughter does (a diagnostic inference). This is a counternormative result and thus inconsistent with causal Bayes net theory on the plausible assumption that the base rate probability of blue eyes is the same across generations. Along the same lines, people were more confident in judging someone's weight from their height than vice versa.

Several other phenomena suggest that people use causal information in nonnormative ways, in contrast to the causal Bayes nets view. Tversky and Kahneman (1983) reported conjunction fallacies that arose when people judged the likelihood of the conjunction of a cause and its effect relative to the likelihood of the effect alone, ostensibly because of undue focus on the causal relation and neglect of the base rate of the cause. Similarly, subadditivity in probability judgments arises when events are unpacked into their typical causes (Tversky & Koehler, 1994), but superadditivity often arises when they are unpacked into atypical causes (Sloman, Rottenstreich, Wisniewski, Hadjichristidis, & Fox, 2004). These phenomena imply that when people have a simple causal relation to focus on, they neglect to search for and use other relevant information such as base rates or alternative causes. Another example is that people judge counterfactual outcomes more likely when the causal chain of events leading to them is easier to simulate regardless of probability, suggesting that people use the ease of causal simulation as a heuristic for probability (Galinsky & Moskowitz, 2000; Kahneman & Tversky, 1982; Wells & Gavanski, 1989). All of these phenomena are inconsistent with basic laws of probability theory and therefore inconsistent with the causal Bayes nets view. In fact, Gilovich and Griffin (2002) listed causality among the six fundamental heuristics of human cognition in a recent review of the judgment literature.

The question of the relation between probability and causal beliefs is also central to other areas of cognition. Studies in the *category-based induction tradition* (Rips, 1975) are inspired by the idea that the prototypical example of inductive inference is the projection of a property from one category to another (Goodman, 1955). For instance, learning that a lion has a property increases the probability that a tiger also has the property. Many approaches to modeling such inferences are based on the similarity between categories (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Sloman, 1993). However, the weight of evidence suggests that similarity is insufficient to capture inferences, especially when people have knowledge about the property being projected (for a review, see Rips, 2001). In particular, causal beliefs are crucial to people's inferences (Heit & Rubinstein, 1994).

Medin, Coley, Storms, and Hayes (2003) explored the effect of causal directionality on category-based inferences using a manipulation analogous to that of Tversky and Kahneman (1980). By varying whether the evidential category was a predator and the conclusion category was prey or vice versa, they manipulated whether participants were asked for a predictive or diagnostic judgment. Like Tversky and Kahneman, they found that predictive inferences were judged higher than diagnostic inferences, a phenomenon they referred to as *causal asymmetry*. For instance, the likelihood of lions having a property given that gazelles have that property was judged to be higher than the likelihood of gazelles having a property given that lions do because there is a relation of transmission from gazelles to lions through the food chain. Like Tversky and Kahneman, they attributed the asymmetry to the greater ease of reasoning from cause to effect than effect to cause. However, it is unclear whether appealing to such nonnormative considerations is necessary to explain the asymmetry. Causal Bayes nets may account for Medin et al.'s results on the assumption that the materials satisfy certain conditions (Shafto, Kemp, Baraff Bonawitz, Coley, & Tenenbaum, 2008).

Current Goals

Our objective is to evaluate whether and to what extent probability judgments conform to a Bayesian model based on causal structure. The basic method is to vary causal structure and to evaluate how predictive and diagnostic probability judgments change relative to the normative standard.

Our experiments have two components: one quantitative and one qualitative. The quantitative component was that in addition to predictive and diagnostic judgments, we collected judgments of people's underlying causal beliefs about the scenarios. We then computed consistent probability judgments on the basis of those beliefs from a normative model and compared the probability judgments to this standard. This allowed us to assess consistency

tapraid5/zfr-xge/zfr-xge/zfr00111/zfr2190d11z | xppws | S=1 | 12/7/10 | 6:12 | Art: 2009-0419

for a wide range of scenarios and to correct for differences in underlying beliefs across participants.

The qualitative component was a manipulation of the strength of alternative causes, because it highlighted the distinction between predictive and diagnostic judgments. Alternative causes should weaken diagnostic judgments because they increase the likelihood that the effect was brought about by a different mechanism. They should also increase predictive judgments for the same reason. To illustrate, consider the predictive and diagnostic questions in (a) and (b):

(a) A mother has a drug addiction. How likely is it that her newborn baby has a drug addiction?

(b) A newborn baby has a drug addiction. How likely is it that the baby's mother has a drug addiction?

Here, the alternative causes are weak; there are relatively few ways a baby can become drug addicted aside from the mother's drug addiction. The causal Bayes net view predicts that people should be sensitive to this factor in the appropriate ways. In (a), the judged probability should reflect the strength of the mother's drug addiction for the effect, whereas in (b), the absence of alternative causes predicts that the effect is highly diagnostic of the cause: if the baby is addicted to a drug, then the mother must be addicted. These considerations might lead to a reversal of the pattern found by Medin et al. (2003) and Tversky and Kahneman (1980), with the diagnostic direction being judged stronger. The question of the extent to which participants consider alternatives is of particular importance because neglect of alternatives is common in human cognition. We return to this point in the General Discussion.

Scope of Work

Not all judgments of probability have their source in causal beliefs. Some arise from the use of noncausal heuristics like anchoring and adjustment (Tversky & Kahneman, 1974) or availability (Tversky & Kahneman, 1973), and some arise from naïve extensional reasoning (Fox & Levav, 2004; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). Nevertheless, we do believe that the cognitive system gives priority to causal relations when they are available and treats causal relations as generating probabilistic ones. As a result, people are more likely to remember causal explanations than the data on which the explanations are based (Brem & Rips, 2000) and are more likely to make predictions that are based on causal assumptions than on probabilistic data (Chapman & Chapman, 1969). Therefore, though our analysis does not bear on every phenomenon in the psychology of judgment, it is relevant to some of the most important ones, and we aimed to add to the growing literature demonstrating the importance of causal structure in understanding a broad range of probability judgments.

Like Tversky and Kahneman (1980), our focus was on comparing predictive to diagnostic inferences. Predictive and diagnostic relations are not a comprehensive set of the possible causal relations between evidence and hypothesis. For instance, evidence and hypothesis can both be correlated effects of a common cause as in a case in which observing that someone has yellow teeth increases the judged likelihood they will get cancer (due to an increased likelihood they smoke). Nevertheless, the contrast between predictive and diagnostic inference is informative because the two relations are primitive in the sense that all other connected causal structures can be reduced to combinations of predictive and diagnostic relations.

Finally, as in the drug-addiction example given earlier and as in Tversky and Kahneman's (1980) "blue eyes" example, we constructed scenarios in which a property was transmitted from one category to another and asked participants to infer whether the effect category would have the property given that the cause category does and vice versa. We did this for three reasons. First, several quantitative models have been advanced to account for category-based induction, and designing our stimuli in this way allowed us to test whether existing models can account for the results. Second, transmission is a very general concept, argued by some philosophers to be the defining feature of what makes a relation causal (e.g. Dowe, 2000). This generality was exemplified by our stimulus set, which included a broad range of causal scenarios, ranging from drug addiction to food preparation, politics, automobiles, and more. Finally, the design gave us a form of control over the stimuli. We were able to manipulate predictive versus diagnostic inference by simply asking for the converse conditional probability, and we were able to manipulate strength of alternative by changing the transmitted property. The norm in the category-based induction literature is a very narrow scope, usually inferences about "blank" predicates applied to a small number of animal categories. In contrast, typical studies in the judgment literature are based on a small number of less well-controlled items. The middle ground between scope and control that we aimed for allowed us to combine some of the virtues of both approaches.

Normative Analysis of Predictive and Diagnostic Judgment

In this section, we use the causal Bayes net framework to develop a normative model of how predictive and diagnostic judgments should change as a function of the underlying beliefs about a causal scenario. Throughout the article, we formalize both kinds of inferences as conditional probabilities. A predictive judgment, which we refer to as *P,* is intended to be an estimate of $P(Effect|Cause)$ while a diagnostic judgment, *D,* is an estimate of $P(Cause|Effect)$.

A critical determinant of these probabilities is the strength of the alternative causes. Alternative causes should weaken *D* because they increase the likelihood that the effect was brought about by a different mechanism, but they should strengthen *P* because they increase the probability that another mechanism brings about the effect even if the cause fails to do so. Another important determinant of predictive and diagnostic judgments is the probability that the cause is effective in bringing about the effect when it is present, what Cheng (1997) called *causal power.* A strong cause is more likely to bring about the effect and hence should yield higher predictive judgments. For the same reason, it should also yield higher diagnostic judgments. A third factor is the prior probability of the cause in question, which should affect only diagnostic judgments. For instance, rare causes should yield low diagnostic judgments, all else being equal, because they are unlikely to have occurred. Predictive judgments should be independent of the prior

probability of the cause because they should reflect only cases in which the cause was present.

Our goal in the normative analysis was to capture the contribution of alternative causes, causal power, and prior probability to predictive and diagnostic reasoning in a way that is probabilistically coherent. A transmission argument can be represented by a common-effect structure, one effect with multiple possible causes. In general, a predicate might be transmitted to the effect category from the target cause or from some alternative generative cause. To capture the additional constraint that a true alternative cause should be independent of the target cause, we restricted ourselves to arguments in which transmission from a source to a recipient follows an independent causal path. Kelley (1972) proposed the *multiple sufficient causes* schema to describe independent causes that combine to generate an effect according to an inclusive-or function. Any of the causes is individually sufficient to bring about the effect, and if more than one cause is present, the effect is also present. In real-world scenarios, a cause is not often strictly sufficient to generate an effect because other things may happen to disable or prevent the effect from happening. This can be modeled by the probabilistic extension of the inclusive-or, sometimes called the *noisy-or*, function. The presence of either cause raises the probability of the effect, and if both causes are present, the probability of the effect is even higher, increasing according to the independent contribution of each cause. When the noisy-or model applies, the calculations of *P* and *D* specified by the model are the only ones that are consistent with the parameters. In that sense, the model offers a normative benchmark for arguments that concern an appropriate causal model. We chose arguments to satisfy the necessary conditions: target and alternative causes were each sufficient for the effect and as independent from each other as possible.

Model Description

A causal Bayes net can be fully described by the probability distributions of its exogenous variables (i.e., variables that have no parents in the graph) along with a set of functions and parameters that define the probability distributions of endogenous nodes conditioned on their parents. In other words, the model requires specification of the prior probability distributions of all root causes and functions describing how causes combine to generate effects.

By aggregating all alternative generative causes into a single node (i.e., a causal background; Cheng, 1997) and aggregating all enablers and disablers into the conditional probability functions, one can concisely represent the structure necessary for defining *P* and *D* as a causal Bayes net with three nodes: the cause, the effect, and the aggregate of all alternative causes. Separate edges connect the cause and alternative to the effect. To specify the parameters over this structure, we assumed that events are binary; they either happen or they do not. This allowed us to represent the probability distribution of exogenous nodes with a single number, a prior probability. We also assumed that the cause and any alternative causes are independent and generate the effect independently according to a noisy-or function as discussed earlier. The independent contribution of a cause can be defined in the model as a parameter that specifies the conditional probability of the effect, given that cause and no other generative causes (a causal power). Disablers are not represented by nodes in the model but instead determine the probabilities with which generative causes bring about the effect. Because of its use of the noisy-or function and parameterization in terms of causal powers, the structure is identical to that proposed in Cheng's seminal power theory of the probabilistic contrast model (PowerPC model) of causal learning.[1] Fn1

To simplify calculations, we collapsed the prior probabilities and causal powers of the alternative causes into a single parameter denoting the strength of alternatives, set to $P(Effect|\sim Cause)$. This is akin to setting their prior probability to 1 (i.e., assuming alternatives are always present but only effective in bringing about the effect some of the time). The prior probability and causal power of alternatives are always confounded in the model, so the simplification is not substantive.

The model is therefore fully parameterized by three numbers: the prior probability of the cause (P_c), the causal power of the cause (W_c) equal to $P(Effect|Cause, \sim Effective\ Alternative\ Causes)$, and the strength of alternatives (W_a) or $P(Effect|\sim Cause)$. The structure and parameterization are depicted in Figure 1. In the figure, W_a represents both the prior and causal powers of alternatives collapsed into a single term. Disablers are implicit in the parameters W_c and W_a. F1

The P and D correspond to $P(Effect|Cause)$ and $P(Cause|Effect)$, respectively. P is calculated with the noisy-or equation:

$$P = P(Effect|Cause) = W_c + W_a - W_cW_a \quad (1)$$

Note the difference between W_c and P. The predictive judgment, P, represents the probability that the effect occurs given that the cause occurred. This includes both cases in which the cause was effective in generating the effect and cases in which the cause was not effective but an alternative cause was. Therefore, P is higher than W_c and increases with the strength of alternatives.

The diagnostic judgment, D, is derived by consideration of its complement, the probability that the cause did not occur despite the effect having occurred (for an alternative derivation, see Waldmann, Cheng, Hagmayer & Blaisdell, 2008).

$$D = P(Cause|Effect) = 1 - P(\sim Cause|Effect) \quad (2)$$

By Bayes' rule:

$$D = 1 - P(Effect|\sim Cause)$$

$$\frac{P(\sim Cause)}{P(Effect)} = 1 - p(\sim Cause)\frac{P(Effect|\sim Cause)}{P(Effect)} \quad (3)$$

Deriving $P(Effect)$ by the noisy-or equation and substituting W_a for $P(Effect|\sim Cause)$ and $(1 - P_c)$ for $P(\sim Cause)$:

$$D = 1 - (1 - P_c)\frac{W_a}{P_cW_c + W_a - P_cW_cW_a} \quad (4)$$

Equation 4 shows that two factors determine D, the prior probability of the cause and the probability that the alternatives caused

[1] According to the PowerPC model of causal learning, causal powers are inferred from contingency data on the assumption that causes contribute to effects independently (i.e., according to a noisy-or model). Our model captures inference rather than learning. Causal power is given, and conditional likelihoods of causes and effects are inferred.

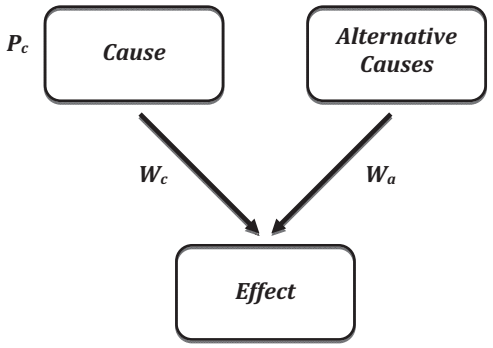

Figure 1. A Bayes net model of transmission arguments. P_c represents the prior probability of the cause, W_c is the causal power of the cause, and W_a is the strength of alternatives, the aggregate causal power and prior probabilities of all alternative causes collapsed into a single term. The effect is generated by a noisy-or function of the cause and the alternatives.

the effect, the ratio between W_a and the extension of $P(Effect)$ at the end of Equation 4. The presence of the effect cannot decrease the probability of the cause, so D is always higher than P_c, and it increases with P_c. Conversely, the effect is diagnostic of the cause to the extent it was not generated by alternative causes. Therefore, the cause and the alternatives compete as the explanation of the effect, and D decreases with the probability that the alternative causes caused the effect.

Model Predictions

Equations 1 and 4 yield predictions regarding how judgments of P and D should vary as a function of the parameters P_c, W_c, and W_a. P is a function of two parameters, W_c and W_a, and increases as each of them increases independently. D is a more complex function of all three parameters; it depends on the prior probability of the cause and the probability that the effect was caused by the alternatives. The probability that the effect was caused by the alternatives is a comparative measure of the strength of alternatives relative to the strength of the cause. Accordingly, it increases with W_a and decreases with P_c and W_c. Therefore, D increases as P_c or W_c increases or as W_a decreases.

Experiment 1

In Experiment 1, we compared predictive and diagnostic judgments about arguments in which there are either strong or weak alternative causes, and we manipulated the strength of alternatives by keeping the categories constant while varying the predicate. Alternative causes, prior probability, and causal power were never mentioned explicitly so in the experiment, we tested people's ability to use aspects of their intuitive causal models in generating likelihood judgments. According to the normative analysis, all else being equal, P should increase with strong alternatives, while D should decrease. If people neglect alternative causes, then varying the strength of alternatives should have little effect on P or D.

Our ultimate goal in Experiment 1 was to generate enough data to test whether the normative model could account for people's predictive and diagnostic judgments. We therefore collected judgments of the model parameters P_c, W_c, and W_a along with predictive and diagnostic judgments. If people's inductive judgments are consistent with their beliefs about the relevant probabilities, then the conditional probabilities derived from the parameters according to the model should match the predictive and diagnostic judgments.

We relied on pre-existing causal beliefs rather than train people on novel causal systems (e.g. Rehder, 2006). One of the most impressive aspects of cognition is people's ability to consider an enormous quantity and variety of prior knowledge and experience when making judgments. The only way to study how people do that is by asking them about events for which all that knowledge and experience can be brought to bear.

Collecting all of the parameters and fitting the model alleviates some of the concerns associated with using naturalistic materials. Ideally, the items would not vary systematically in the other parameters across the manipulation, and we used a large number of items to try to make that likely. Nonetheless, with naturalistic materials, potential confounding is always a concern. The model fitting allowed us to interpret the results, even in the case of confounding. Thus, the effects across conditions are only suggestive. It is the modeling that provides the real interpretive power.

Another concern with using people's pre-existing beliefs is that we could not be certain how well those beliefs would conform to the model assumptions. The primary concern is that the main cause and the alternative causes might not be completely independent. This is a valid concern but is mitigated by the fact that alternative causes necessarily raise probability. Therefore, even if dependence is introduced, the normative value for P is still higher than W_c, unless the causes are perfectly correlated. We chose materials with the independence assumption in mind, so on average the value for P should be close to the model prediction. An analogous argument applies to judgments of D.

Method

Participants. We recruited 162 participants from college message boards on the Internet; these participants logged on to complete the survey remotely for a chance to win a $100 lottery prize. Additionally, 18 Brown University students were recruited from the psychology research pool or through flyers posted on campus; these participants completed the questionnaire on a computer in our lab to receive either class credit or $8 per hour. In total, 180 participants completed the experiment. The survey was designed and administered through the SurveyMonkey service, and the survey software ensured a single response per computer.

Design. The experiment had three independent variables: categories, strong versus weak alternatives, and question type. Categories and predicates were chosen to fit the common effect noisy-or causal structure in which any alternative causes provide an independent contribution to the effect and the causal relation from cause to effect is unidirectional. For each predicate, we asked five questions: the prior probability of the cause (P_c), the causal power of the cause (W_c), the strength of alternatives (W_a), the predictive judgment (P), and the diagnostic judgment (D). To probe for these, we asked for the likelihood of the relevant events rated on a 0–100 scale. Examples of the wordings of the questions for one item are shown in Table 1. We chose 20 sets of categories, T1

tapraid5/zfr-xge/zfr-xge/zfr00111/zfr2190d11z | xppws | S=1 | 12/7/10 | 6:12 | Art: 2009-0419

Table 1
Example Question Forms From Experiment 1

Parameter/judgment	Wording of example
Prior probability of cause (P_c)	A woman is the mother of a newborn baby. How likely is it that the woman is drug addicted?
Causal power of cause (W_c)	The mother of a newborn baby is drug addicted. How likely is it that her being drug addicted causes her baby to be drug addicted?
Strength of alternatives (W_a)	The mother of a newborn baby is not drug addicted. How likely is it that her baby is drug addicted?
Predictive judgment (P)	The mother of a newborn baby is drug addicted. How likely is it that her baby is drug addicted?
Diagnostic judgment (D)	A newborn baby is drug addicted. How likely is it that its mother is drug addicted?

two predicates for each set, and five questions for each predicate for a total of 200 questions. The predicates and categories are shown in Appendix A.[2]

To avoid interactions among questions about the same predicate, we assigned the 200 questions to one of five questionnaires of 40 questions each. Each participant received one questionnaire. Questions were randomly assigned with the constraints that each questionnaire had one question type from each of the 40 predicates and that no questionnaire had the same question type of the weak and strong predicate for a given set of categories. Each participant therefore answered a single question about each predicate. The order of questions in each questionnaire was randomized but constant for each questionnaire.

Materials and procedure. Participants were randomly assigned to receive one of the five questionnaires. Each questionnaire consisted of instructions at the top followed by 40 questions, all on a single screen. Participants were instructed to "[g]ive an answer between 0 (impossible) and 100 (definite)" for each question. The experiment took approximately 20 min.

Results

Five participants gave the same response to all 40 questions and were omitted from subsequent analysis. The mean predictive and diagnostic judgments for the strong and weak alternatives conditions are shown in Figure 2. We collapsed the data across participants and assessed the relative effect of strength of alternatives on predictive and diagnostic judgments by performing a 2 (alternatives: strong vs. weak) × 2 (judgment: predictive vs. diagnostic) repeated-measures analysis of variance (ANOVA).[3] There was a significant interaction between judgment type and strength of alternatives, $F(1, 19) = 31.4$, $p < .001$, partial $\eta^2 = .62$. There was also a main effect of strength of alternatives, $F(1, 19) = 4.9$, $p = .039$, partial $\eta^2 = .21$, but no significant effect of type of judgment, $F(1, 19) = 0.6$, *ns*.

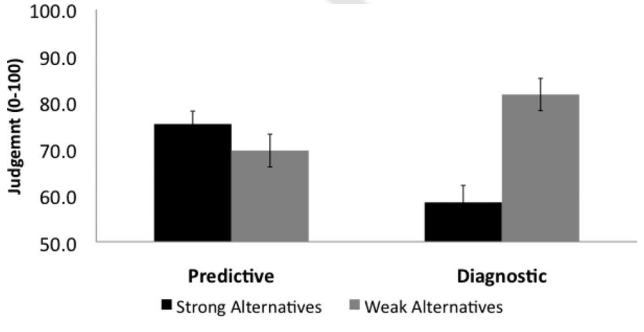

Figure 2. Mean predictive and diagnostic judgments and standard errors for the strong and weak alternatives conditions of Experiment 1.

We conducted planned comparisons between judgments in the strong and weak alternatives conditions. Diagnostic judgments in the weak alternatives condition ($M = 81.7$) were higher than in the strong alternatives condition ($M = 58.5$), $t(19) = 5.0$, $p < .001$, Cohen's $d = 1.1$. Predictive judgments did not differ significantly ($M_{strong} = 75.3$; $M_{weak} = 69.6$) $t(19) = 1.3$, *ns*. Corroborating this analysis, we also found that there was no significant difference between judgments of P and W_c, $t(39) = 0.60$, *ns*.

We also used matched sample t tests to compare mean parameter judgments for each category set across the strong/weak manipulation. The results are shown in Table 2. W_a judgments were higher in the strong alternatives condition than in the weak alternatives condition ($p < 0.001$), validating the experimental manipulation. P_c and W_c responses did not differ significantly between conditions.

Model fits.

Modeling details. The model represents the relation between a single participant's judgments of the parameters P_c, W_c, and W_a and their judgments of P and D. Because of the incomplete design, no participant made all of the parameter judgments for any single item, and we therefore had a distribution of unmatched judgments of the parameters for each item. We could not simply take the means of these distributions and combine them according to the model's equations because it is not generally true that the mean of a function of distributions is equivalent to the application of that function to their means. In particular, the equation for D, which includes random variables in the denominator, violates this assumption. For P, the assumption did hold, and the model's outputs for P were the same as if they were calculated directly from the parameter means. Nonetheless, for consistency's sake, we used the same procedure to generate predictions for P and D.

[2] Some of the categories could be described as having part–whole relationships, but we still consider them transmission scenarios because the predicate applies to the part before it applies to the whole. The predicates were such that if the predicate applied to the part, it increased the probability that the predicate applied to the whole but did not make it necessary. Therefore, the causal structures of these items do not differ from the rest.

[3] Each participant did not make the same number of judgments for each dependent variable; thus, not all participants supplied a sufficient number of judgments per condition to support an analysis by participants. We therefore collapsed over participants and used the category means for all of the analyses of Experiment 1.

Table 2
Mean Parameter Judgments for the Strong and Weak Alternatives Conditions of Experiment 1

Parameter	Strong alternatives	Weak alternatives	*t*	*df*	*p*
Prior probability of cause (P_c)	41.6	48.2	1.14	19	0.27
Causal power of cause (W_c)	75.0	71.4	0.79	19	0.44
Strength of alternatives (W_a)	39.0	20.0	5.00	19	<0.001

Our method was to use a sampling procedure to generate a distribution for the model's predictions of *P* and *D* for each item and used the mean of this distribution as the model's prediction for that item. To generate a single sample of *P* and *D* for a given item, we drew one sample of each of the three parameters uniformly and independently from the set of participant responses. We then calculated *P* and *D* from the sampled parameters according to Equations 1 and 4. We repeated this procedure to generate 100,000 samples each of *P* and *D* for each item and took the means as the model's predictions for that item. Reruns of the sampling procedure yielded no differences in the predictions for either *P* or *D*.

F3 ***Modeling results.*** Figure 3 shows the model predictions for *P* (left panel) and *D* compared with participant responses. As with participant responses, model predictions for *D* were higher in the weak condition ($M = 78.6$) than in the strong condition ($M = 61.2$), $t(19) = 5.0$, $p < .001$. Model predictions for *P* were lower in the weak condition ($M = 76.8$) than in the strong condition ($M = 85.3$), $t(19) = 2.38$, $p = .028$, Cohen's $d = 0.5$. The model predictions of *D* were not significantly different from participant responses, $t(39) = 0.7$, *ns*, and were highly correlated with items in the strong and weak conditions separately, $r_{strong} = .69$, $p < .001$; $r_{weak} = .69$, $p < .001$, and across both conditions, $r = .80$, $p < .001$. Model predictions of *P* ($M = 81.1$) were significantly higher than participant responses ($M = 72.5$), $t(39) = 6.54$, $p < .001$, Cohen's $d = 1.09$, but were still highly correlated both within each condition, $r_{weak} = .83$, $p < .001$; $r_{strong} = .75$, $p < .001$, and across conditions, $r = .72$, $p < .001$.

A possible concern is that the normative model is superfluous and that one of the parameters alone can predict judgments of *P* and *D*. We therefore used hierarchical multiple regression analyses to test whether the normative model does better than individual parameters at accounting for the variance in *P* and *D* judgments across items. The results of those analyses are shown in Table 3. T3 For judgments of *D*, we considered the possibility that the high correlation between the model and judgments of *D* could be driven primarily by differences in W_a. We found that W_a was significantly correlated with *D* across the strong/weak manipulation, $r = -.49$, $p = .003$; however, the correlations were not significant in each condition separately, $r_{weak} = -.28$; $r_{strong} = -.08$. The hierarchical multiple regression, in which W_a and the normative model were used as predictors of *D,* showed that the model fit the data better than W_a alone, and W_a had no predictive value beyond its role in the model. Together, the normative model and W_a accounted for 64% of the variance in *D*. The unique variance of the normative model accounted for 41% of the variance of *D*, $F(1, 39) = 41.7$, $p < .001$, but the unique variance of W_a did not account for any of the variance of *D*, $F(1, 39) = 1.0$, *ns.*

In contrast, the best predictor of predictive judgments was the single parameter W_c and not the full model. W_c alone fit the data better than the model, and the model had no predictive value beyond that of W_c. The model and W_c together accounted for 77% of the variance of *P,* most of which was shared variance (67%). The unique variance of W_c accounted for 10% of the variance of *P*, $F(1, 39) = 17.1$, $p < .001$, but the unique variance of the model did not account for any of the variance of *P,* $F(1, 39) = 0.4$, *ns*. In other words, the correlation between the full model and predictive judgments is artifactual and fully modulated by W_c. Because W_c and W_a are the only two factors in the model prediction of *P*, these results imply that predictive judgments were uncorrelated with W_a, which we verified, $r = .04$, *ns*.

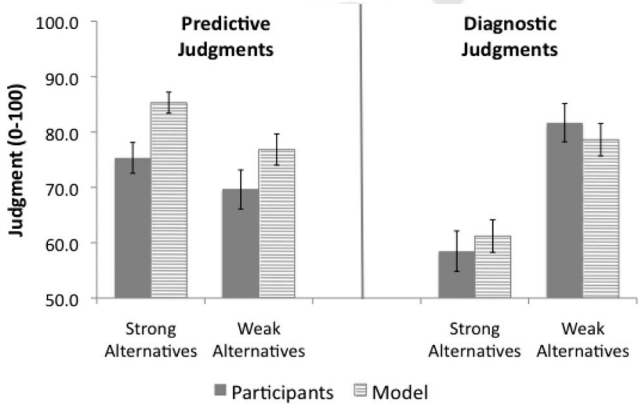

Figure 3. Comparisons between mean participant responses and model predictions for Experiment 1 with standard errors. Predictive judgments are shown in the left panel and diagnostic judgments on the right.

Discussion

Participants were sensitive to alternative strength when reasoning diagnostically but not predictively. We found a large difference of alternative strength for diagnostic judgments but no difference for predictive judgments, despite participants' judging the alterna-

Table 3
Variance of Predictive and Diagnostic Judgments Accounted for by the Normative Model Versus a Single Predictive Parameter

Predictor	All variance	Unique variance	*p*
Diagnostic judgments			
Strength of alternatives (W_a)	0.23	0.01	0.32
Model prediction for *D*	0.63	0.41	<0.001
Predictive judgments			
Causal power (W_c)	0.77	0.10	<0.001
Model prediction for *P*	0.67	0.002	0.53

tapraid5/zfr-xge/zfr-xge/zfr00111/zfr2190d11z | xppws | S=1 | 12/7/10 | 6:12 | Art: 2009-0419

tives twice as strong in the strong condition relative to the weak condition. The model fitting allowed us to rule out possible alternative explanations for this pattern. When we extrapolated predictive judgments using the model, the results were significantly underestimated by the predictive judgments that were probed directly. This underestimation was driven by the lack of consideration of W_a. Predictive judgments were invariant to W_a and were similar to W_c, judgments of causal power.

The model achieved good fits to participants' diagnostic judgments, with zero free parameters. The model did not just achieve a good fit when the data were aggregated over arguments. Instead, the model accounted for a large part of the variance across specific arguments. The model's good fit did not simply capture participants' sensitivity to alternative strength. The model was highly correlated with participant judgments within the strong and weak conditions separately, while W_a was uncorrelated with those judgments and W_a had no predictive value beyond its role in the model. In other words, the strength of alternatives was only important in the context of the other parameters. On average, participants combined information about prior probability, causal power, and alternative strength in a way that approximated the normative computation fairly closely.

Experiment 2

The partially between-participants design of the model fitting in Experiment 1 required us to generate samples from the posterior distribution of D as opposed to our calculating model fits directly from parameters given by participants. In Experiment 2, we sought to replicate the findings of Experiment 1 with a design that allowed us to calculate predicted values of P and D directly from each participant's judgments. This meant that all parameter estimates had to be collected from each participant. We also collected judgments of $P(Effect)$ or P_e (e.g., "A woman is the mother of a newborn baby. How likely is it that the newborn is drug-addicted?"). This allowed us to compare the model with an alternative model of categorical induction.

Method

Participants. We recruited 78 participants by Internet advertisement; they participated online for a chance to win a $100 lottery prize. Additionally, 30 Brown University students participated in the lab to receive either class credit or $8 per hour. In total, 108 participants completed the experiment.

Design. We chose five sets of categories from the 20 that were used in Experiment 1. As in Experiment 1, we also manipulated strong versus weak alternatives by choosing two different predicates for each set of categories, and question type was a third independent variable (we collected judgments of P_c, W_c, W_a, P, D, and P_e). There were five category sets, two predicates for each category and six questions for each predicate for a total of 60 questions. All variables were manipulated within participant, so each participant answered all 60 questions.

To attenuate interactions among items, we split the questions onto three pages so that each predicate was represented in two questions per page, W_c and D, P_c and P, or W_a and P_e. The order of questions on each page was randomized. To test for order effects, we created two versions of the questionnaire. In the second version, the questions and pages were displayed in reverse order from that of the first version. Participants were randomly assigned to one of the two versions.

Procedure and stimuli. We chose five of the category sets from Experiment 1: mother/baby, apple slices/apple pie, football coach/team, engine/Honda Accord, and music/party. The questions were the same as in Experiment 1 except that the wording of the diagnostic question for the weak alternatives coach/team predicate was changed to a more natural form. We also changed the strong alternatives predicate for the engine/Honda Accord question because we were concerned that the statement that the engine was not functioning properly implied that the car did not function properly. We therefore used the predicate "is noisy" instead.

The procedure was identical to that in Experiment 1 except that there were 60 questions instead of 40, and they covered three pages rather than one. We also added the following to the instructions: "Please answer the questions in order. Once you've answered a question, don't go back and change it. Though some of the questions are similar to previous questions, it is important to answer every question in the set." The questionnaire took approximately 30 min to complete.

Results

Judgments. One participant gave the same response to each question and was omitted from subsequent analyses. Responses to the two question orders were highly similar, $r = .98$, $p < .001$. The responses of Internet and lab participants were also highly similar, $r = .98$, $p < .001$. For all subsequent analyses, therefore, we used the full data set collapsed over orders and Internet/lab populations.

The mean predictive and diagnostic judgments for the strong and weak alternatives conditions are shown in Figure 4. We subjected the participant means to a 2 (alternatives: strong vs. weak) × 2 (judgment: predictive vs. diagnostic) repeated-measure ANOVA. Once again, we observed a significant interaction between alternative strength and judgment type, $F(1, 106) = 137.7$, $p < .001$, partial $\eta^2 = 0.6$. There was also a main effect of alternative strength, $F(1, 106) = 6.3$, $p = .014$, partial $\eta^2 = 0.06$, but no main effect of question type, $F(1, 106) = 0.72$, *ns*.

Planned comparisons between judgments in the strong and weak alternatives conditions revealed that diagnostic judgments in the weak alternatives condition ($M = 83.5$) were higher than in the strong alternatives condition ($M = 70.4$), $t(106) = 9.1$, $p = <0.001$, Cohen's $d = 0.9$. Unlike Experiment 1, predictive judg-

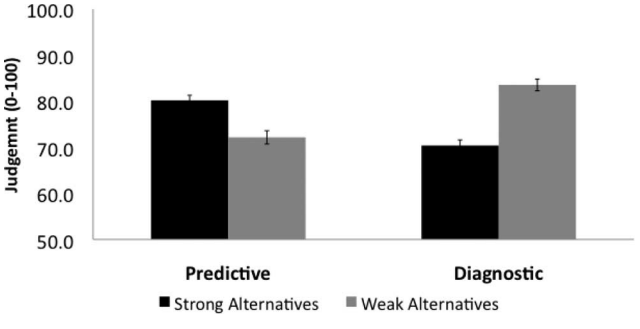

Figure 4. Mean predictive and diagnostic judgments and standard errors for the strong and weak alternatives conditions of Experiment 2.

ments were significantly higher for strong items than weak ones, (M_{strong} = 80.2; M_{weak} = 72.2), $t(106) = 6.3$, $p < .001$, Cohen's $d = 0.6$.

To assess any parameter differences across the strong/weak manipulation, we performed matched-sample t tests on question means (Table 4). Replicating Experiment 2, W_a was judged higher for the strong items than the weak ones. P_c, W_c and P_e were also judged higher for strong items than weak ones. Due to the parameter differences between conditions, no conclusions about the relative neglect of alternatives for predictive and diagnostic judgments could be drawn without model fitting.

Model fits. Because of the within-participants design we were able to calculate model predictions for each participant and each item instead of sampling as we did in the analysis of Experiment 1. For each participant, we simply took the parameters they gave for a particular item and calculated Equations 1 and 4 to yield a prediction for P and D for that item.

Figure 5 shows model predictions compared with participant responses. As in Experiment 1, the model overestimated participants' predictive judgments in both the strong $t(106) = 9.6$, $p < .001$, Cohen's $d = 1.0$, and weak conditions, $t(106) = 6.4$, $p < .001$, Cohen's $d = 0.4$. The model fits for diagnostic inferences were much closer. In the strong condition model, predictions and participant judgments were not significantly different, $t(106) = 1.3$, *ns*. In the weak condition, participant responses were lower than the model estimates, but this difference was very small, $t(106) = 2.1$, $p = .04$, Cohen's $d = 0.2$.

A hierarchical multiple regression analysis over individual participants' responses also showed the same pattern as in Experiment 1. The variance of diagnostic judgments was better accounted for by the model than by W_a, and the variance of predictive judgments was better accounted for by W_c than by the model. Once again, W_a was uncorrelated with predictive judgments, $r = .03$, *ns*.

Discussion

The results of Experiment 2 corroborated the conclusions of Experiment 1. The pattern of results was somewhat different than in Experiment 1 as predictive judgments were significantly higher in the strong alternatives condition than the weak alternatives condition. However, the model fitting showed that this difference was not due to differences in W_a. Once again, predictive judgments were invariant to alternative strength and were lower than the predictions of the model. The differential pattern from Experiment 1 was likely due to the small number of categories used in the experiment. Also corroborating Experiment 1, the model predicted diagnostic judgments more accurately than predictive judgments.

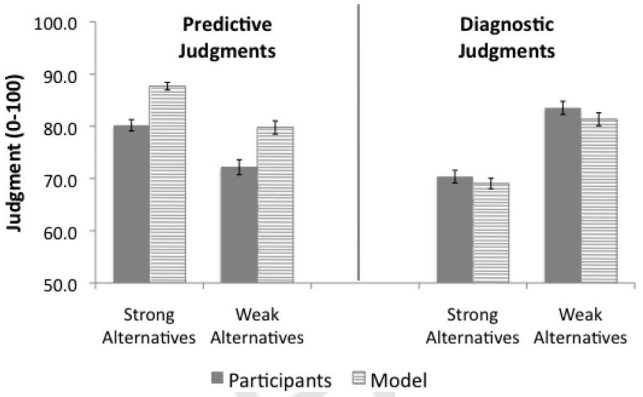

Figure 5. Comparisons between mean participant responses and model predictions for Experiment 2 with standard errors. Predictive judgments are shown in the left panel and diagnostic judgments on the right.

Experiment 3

The conclusion from Experiments 1 and 2 that participants neglected alternatives in the predictive direction was based in part on the similarity between predictive judgments and judgments of causal power, W_c. We attributed this to how people reason, but it could instead reflect how they interpreted the questions. One possibility is that participants may have interpreted the W_c question as asking for P. In the W_c question, participants are asked to judge the likelihood that the cause causes the effect. Participants might not understand this question as asking for causal power and give a conditional probability judgment instead.

In Experiment 3, we tested this possibility by mentioning an alternative cause explicitly and then asking the W_c and P questions. We expected participants to take the mentioned alternative into account and give higher P judgments than W_c judgments as in the normative model. The misinterpretation hypothesis predicts that judgments of P and W_c should be the same even when alternatives were mentioned.

Method

Participants. We recruited 62 Brown University students on campus, and they participated voluntarily, with 31 students assigned to each condition.

Design. We chose 10 of the strong alternative items from Experiment 1 to maximize the effect of mentioning the alternative cause. The main independent variable was whether participants were asked for judgments of P or W_c, and it was manipulated between participants. Each participant therefore answered either

Table 4
Mean Parameter Judgments for the Strong and Weak Alternatives Conditions of Experiment 2

Parameter	Strong alternatives	Weak alternatives	t	df	p
Prior probability of cause (P_c)	50.6	42.4	7.57	106	<0.001
Causal power of cause (W_c)	78.5	74.7	2.90	106	0.005
Strength of alternatives (W_a)	38.9	17.2	15.62	106	<0.001
Prior probability of effect (P_e)	49.4	34.6	11.45	106	<0.001

tapraid5/zfr-xge/zfr-xge/zfr00111/zfr2190d11z | xppws | S=1 | 12/7/10 | 6:12 | Art: 2009-0419

10 *P* questions or 10 W_c questions. All of the questions explicitly mentioned the possibility of an alternative cause without saying whether that cause was present. An example of a *P* question is "The coach of a high school football team is highly motivated. Accolades from family and friends could also cause high school football teams to be highly motivated. How likely is it that the team is highly motivated?" The analogous W_c question was "The coach of a high school football team is highly motivated. Accolades from family and friends could also cause high school football teams to be highly motivated. How likely is it that the coach being highly motivated causes his team to be highly motivated?"

Procedure and Stimuli. Participants were handed a single sheet with the 10 questions and instructions at the top. The questionnaire took between 5 and 10 min to complete. The stimuli used in the experiment are shown in Appendix A.

Results

Due to a typographical error in one of the questionnaires, the Honda Accord item was omitted from the analysis. The mean *P* and W_c judgments for Experiment 3 are shown in Figure 6 along with those for the same items from Experiment 1 for comparison. An independent sample *t* test on participant means revealed that judgments of *P* ($M = 81.7$) were significantly higher than W_c ($M = 72.0$), $t(60) = 3.5$, $p < .001$, Cohen's $d = 0.9$, as predicted by the neglect hypothesis. A matched sample *t* test on category means yielded the same result.

Discussion

In Experiment 3, judgments of *P* were higher than W_c when the possibility of an alternative cause was mentioned explicitly. Judgments of W_c were similar to judgments of both *P* and W_c for the same items from Experiment 1. This suggests that participants took alternative causes into account in judging *P* but only when alternatives were mentioned explicitly. The increase in judgments of *P* was not brought about by giving people new information but rather by directing their attention to something they already knew. For example, most participants were likely aware that accolades from family and friends might motivate high school football teams.

These results rule out the possibility that participants are answering the *P* question when they are asked the W_c question. However, there is an additional possibility that participants interpret the *P* question as asking for W_c. Experiment 3 provides evidence against this possibility because participants treated the questions differently, but it cannot be ruled out. The possibility remains that participants understand that they should take into account not only causes mentioned in the question itself but also those mentioned in the context of the question, even if those causes are not definitively present. Thus in Experiment 3, they might have interpreted the *P* questions as asking for the probability of the effect conditioned on the presence of the main cause and the possibility of the alternative cause mentioned, but no other causes, which would have led to higher judgments than for the W_c questions. Extending the pragmatic hypothesis in such a way makes it much more difficult to pin down and differentiate from neglect of alternatives. In Experiment 4, we addressed this extended version of the pragmatic account differently: by reducing the vagueness of the questions we posed.

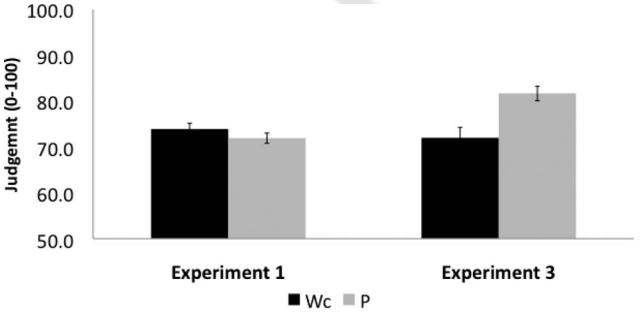

Figure 6. Mean *P* and W_c judgments for Experiment 3, with the judgments for the same items from Experiment 1.

Experiment 4

Hertwig and Gigerenzer (1999) have shown that people have a variety of different interpretations of probability and that questions about frequency are less vague. Therefore, in this experiment, instead of asking participants for the likelihood of the effect given the cause, we specified a definite set of instances and asked participants to estimate the frequency of a subset. The following is an example:

> (a) Consider mothers who each have a single newborn baby. Of 100 mothers who are drug addicted, how many of the mothers' babies are drug addicted?

Participants are asked to estimate the number of babies out of 100 that are drug addicted. To interpret this question as asking for W_c would imply that one should not include drug-addicted babies whose drug addiction is due to some other source besides the mother. This would be an odd interpretation, given that the question explicitly asks for the number of drug-addicted babies. A further benefit of frequency formats is that they are one way to obtain more veridical representations of uncertainty. A number of experiments have shown that clarifying the relations among relevant sets reduces the incidence of fallacies in probability judgment (reviewed in Barbey & Sloman, 2008). One way to make set relations transparent is through the use of frequency formats (e.g., Gigerenzer & Hoffrage, 1995; Tversky & Kahneman, 1983). Experiment 4 thus serves to test the robustness of the inductive asymmetry.

Method

Participants. In this experiment, 68 undergraduates from the Brown University psychology pool participated for class credit.

Design, stimuli, and procedure. We asked three types of questions: *P, D* and W_c. *P* questions were phrased as in Example (a). *D* and W_c questions were phrased as in Examples (b) and (c) respectively:

> (b) Consider mothers who each have a single newborn baby. Of 100 babies who are drug addicted, how many of the babies' mothers are drug addicted?

> (c) Consider mothers who each have a single newborn baby. Of 100 mothers who are drug addicted, in how many cases does the mother being drug-addicted cause her baby to be drug addicted?

tapraid5/zfr-xge/zfr-xge/zfr00111/zfr2190d11z | xppws | S=1 | 12/7/10 | 6:12 | Art: 2009-0419

We utilized all 20 category sets and the strong and weak predicates from Experiment 1. The 120 questions were divided into three questionnaires such that no questionnaire had the strong and weak version for a particular question type. Each participant was assigned at random to one of the three questionnaires and completed the experiment in approximately 20 min.

Results

F7 The results of Experiment 4 are depicted in Figure 7. We collapsed over categories and subjected the data to a 2 (predictive vs. diagnostic) × 2 (strong vs. weak) ANOVA. As in Experiment 1, there was a significant interaction between strength of alternatives and direction of inference, $F(1, 67) = 46.0$, $p < .001$, partial $\eta^2 = 0.4$. There was also a main effect of direction of inference, $F(1.67) = 9.5$, $p = .003$, partial $\eta^2 = 0.4$, and strength of alternatives, $F(1, 67) = 43.4$, $p < .001$, partial $\eta^2 = 0.1$. Collapsing the data over participants and comparing question means yielded a similar pattern: a significant interaction, $F(1, 19) = 19.1$, $p < .001$, partial $\eta^2 = 0.5$, and a main effect of direction of inference, $F(1, 19) = 5.6$, $p = .029$, partial $\eta^2 = 0.2$, but no main effect of strength of alternatives, $F(1, 19) = 2.0$, *ns*.

We performed a series of planned comparisons to test the impact of strength of alternatives. As in Experiment 1, there was a large difference between judgments of *D* across the strong/weak manipulation ($M_{strong} = 69.0$, $M_{weak} = 86.8$), $t(67) = 8.1$, $p < .001$, Cohen's $d = 1.0$, but no difference for judgments of *P* ($M_{strong} = 82.2$, $M_{weak} = 81.5$), $t(67) = 0.5$, *ns*. Judgments of *P* and W_c did not differ for either the strong ($M_{wc} = 79.5$), $t(67) = 1.1$, *ns*, or weak ($M_{wc} = 78.8$), $t(67) = 1.3$, *ns* items. Collapsing the data over participants and comparing question means yielded the same results: a large difference between judgments of *D* across the strong/weak manipulation, $t(19) = 4.9$, $p = .001$, Cohen's $d = 1.1$; no difference for judgments of *P,* $t(19) = 0.6$, *ns*; and no difference between *P* and W_c for either strong, $t(19) = 1.5$, *ns*, or weak, $t(19) = 1.4$, *ns*, predicates.

Discussion

The results of Experiment 4 corroborated Experiments 1 and 2 with less vague frequency-formatted questions. This supports the hypothesis that the failure to consider alternatives in predictive judgment is not driven by participants' interpreting *P* questions as requesting W_c but rather by their neglecting alternative causes. W_c questions were rated as slightly lower than the *P* questions. Though this difference was not significant, one might ask whether it might have become so with additional data. While this is logically possible, the small difference is not sensitive to the strong/weak manipulation, suggesting that it does not represent even partial consideration of alternatives.

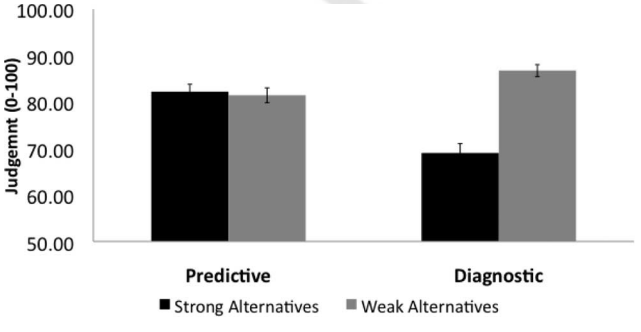

Figure 7. Mean predictive and diagnostic judgments and standard errors for the strong and weak alternatives conditions of Experiment 4.

Experiment 5

Experiment 5 provides an even stronger test of the pragmatic account. Following Bonini, Tentori, and Osherson's (2004) methodology, we removed any pragmatic context from the judgment task by telling people that "an impartial judge who does not know the evidence will check to see if the event has in fact occurred. Your job is to determine how likely it is that the judge will find that the event has in fact occurred." Each conditional question was phrased, "How likely is it that the judge will determine that . . . ." Since the judge was ignorant of the evidence, it would have made no sense for participants to construe the likelihood question as requesting that the judgment be based solely on the cause mentioned in the evidence.

Additionally, we explicitly instructed participants that the mention of particular evidence did not rule out other evidence. We also stressed that the task was to judge the likelihood of events. To this end, we included some items without any evidence, where the task was to judge the marginal likelihood of the event. Given these differences, participants could not have interpreted the predictive questions as requesting causal power. We predicted that the pattern of results would replicate those of Experiments 1, 2 and 4, providing more evidence against the pragmatic account.

Method

We recruited 28 undergraduates from the Brown University psychology pool who participated for class credit. We used the same category sets and predicates as Experiments 1 and 4 and collected judgments of *P* and *D* for a total of 80 questions. The 80 questions were split into two groups such that the *P* and *D* questions for a particular predicate never co-occurred. Participants were assigned at random to one of the two sets of questions. We added 20 filler items to each set of questions; eight were conditional questions, and 12 were marginal questions. In total, each participant answered 60 questions.

The experiment was conducted on a computer in the lab. Participants first read the following:

> In this experiment, you will estimate the likelihood of particular events. Sometimes you will be given some evidence and be asked to judge how likely the event is, given that you know the evidence. Sometimes you will be asked to estimate the likelihood of the event without any evidence.
>
> In both cases, there may be relevant information that is not mentioned in the evidence. Just because something is not mentioned, that does not mean it is absent. It just means you do not know whether it is present or absent.
>
> When determining your answer, imagine that an impartial judge who does not know the evidence will check to see if the event has in fact occurred. Your job is to determine how likely it is that the judge will find that the event has in fact occurred.

tapraid5/zfr-xge/zfr-xge/zfr00111/zfr2190d11z | xppws | S=1 | 12/7/10 | 6:12 | Art: 2009-0419

> Give an answer between 0 (*impossible*) and 100 (*definite*). You will give your answer by typing the number on the keyboard. When you have finished answering, hit the return key to move on to the next question. Try to answer as quickly and as accurately as you can.
>
> Do not think too hard about each question as there is no correct answer, but do not guess wildly either.
>
> Hit the return key to begin the first question.

After reading the instructions, participants proceeded to the questions. The order of questions was randomized for each participant. One question was displayed on the screen at a time. At the top of the screen, the word "Evidence" was displayed to the left of where the evidence appeared. On trials with conditional questions, the evidence was phrased as in Example (a).

> (a) The mother of a newborn baby is drug addicted.

On trials with no evidence, the words "no evidence" were displayed. The likelihood question was displayed 2 in. beneath the evidence and read as in Example (b).

> (b) How likely is it that the judge will determine the baby is drug addicted?

Participants typed their responses into a box at the bottom of the screen and hit "return" to move to the next question. The entire set of questions took 20–30 min to complete.

Results and Discussion

The results were similar to those from Experiment 1 and 4 and are depicted in Figure 8. Collapsing over categories and comparing participant means produced a significant interaction between strength of alternatives and direction of inference, $F(1, 27) = 36.0$, $p < .001$, partial $\eta^2 = 0.6$. There was also a main effect of direction of inference, $F(1.27) = 13.1$, $p = .001$, $\eta^2 = 0.3$, and strength of alternatives, $F(1, 27) = 23.2$, $p < .001$, $\eta^2 = 0.5$. Planned comparisons revealed a large effect of strength of alternatives on diagnostic judgments ($M_{strong} = 62.0$; $M_{weak} = 76.3$), $t(27) = 5.8$, $p < .001$, Cohen's $d = 1.1$, but no effect on predictive judgments ($M_{strong} = 76.0$, $M_{weak} = 74.5$), $t(27) = 1.4$, *ns*. Collapsing over participants and comparing category means yielded the same result: a large effect of alternative strength on diagnostic judgments, $t(19) = 4.1$, $p < .001$, Cohen's $d = 0.9$, but no effect on predictive judgments, $t(19) = 0.4$ *ns*. The results support the conclusion that failure to consider alternative causes in prediction is the result of participants' beliefs and reasoning processes and not their interpretation of the specific experimental demands posed by the question. The demands of the task in this experiment were clearly to give a judgment in which the absence of alternative causes was not assumed; yet alternative causes were neglected as in previous experiments.

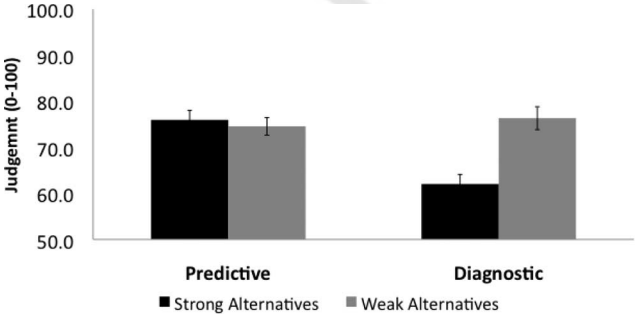

Figure 8. Mean predictive and diagnostic judgments and standard errors for the strong and weak alternatives conditions of Experiment 5.

General Discussion

We have provided a normative analysis of predictive and diagnostic probability judgments and reported five experiments in which we tested how people's inferences compared with the analysis. In Experiments 1 and 2, we collected judgments of causal parameters along with predictive and diagnostic judgments, allowing us to fit a causal Bayes net model. Participants were sensitive to alternative strength when reasoning diagnostically and were also sensitive to the other factors highlighted by the normative analysis—causal power and prior probability—integrating these variables in a ways that closely approximated the model. In the predictive direction, however, the participants neglected alternative causes, leading them to systematically underestimate probability. Experiment 3 provided further evidence for neglect in the predictive direction: mentioning alternatives led to higher *P* judgments. The fact that participants did not raise their W_c judgments provides evidence against pragmatic explanations, suggesting that participants did not differentiate the causal power and conditional probability questions. In Experiment 4, we replicated Experiments 1 and 2 using questions about the number of cases the predicate applied to, providing further evidence that the effect was not due to a misinterpretation of the questions. Experiment 5 provided even stronger evidence against the pragmatic account as no sensitivity to alternatives in prediction was shown when a causal power interpretation of the predictive question was ruled out.

These results highlight both the strengths and weaknesses of causal Bayes net theories. People's probability judgments reflect a sophisticated causal reasoning process that is sensitive to many of the appropriate causal variables. For instance, when making diagnostic judgments, participants integrated multiple causal parameters to measure the relative strength of different causes, as prescribed by the model. This provides some support for Bayesian models of inference that use causal structure when the categories are causally related (e.g. Kemp and Tenenbaum, 2009; Shafto et al., 2008) and those that represent causal structure at the level of individual categories (Rehder, 2009). In contrast, alternative causes were also relevant to predictive judgments, and participants systematically failed to take them into account. Causal Bayes net models do not predict this neglect of alternatives, suggesting that probability judgment cannot be fully explained in a straightforward way by normative models.

We believe that in a limited sense, Tversky and Kahneman (1980) were correct: people find it easier to think from cause to effect. They are simply doing what comes most naturally in that direction, thinking from cause to effect rather than considering the entire relevant causal structure. Alter, Oppenheimer, Epley, and Eyre (2007) have shown that when a question feels easy, people deliberate on it less and make more errors than when the same question feels more difficult. Therefore, the ease of invoking causal power as a response to the predictive question could lead

people to neglect alternatives. Another possibility is that it is precisely the neglect of alternatives that makes predictive questions feel easier than diagnostic ones. These explanations are supported by reaction time data collected by Fernbach and Darlow (2010). Predictive judgments were made faster than diagnostic judgments overall and while reaction time for diagnostic judgments increased with the strength of alternatives, predictive judgments showed no such dependency.

Although we agree with Tversky and Kahneman (1980) regarding the ease of predictive reasoning, our results are inconsistent with a bias to overestimate predictions relative to diagnoses. We found bias in the opposite direction: predictive judgments were too low. Our experiments differed from Tversky and Kahneman's in two fundamental ways: First, they chose situations with identical predictive and diagnostic probabilities. This restriction likely contributed to the small number of examples used to establish the bias phenomenon Our analysis assessed normativity for a wider range of situations because it predicts the relative strength of P and D for all parameter values. This generality allowed us to generate a large set of stimuli. One possibility is that their finding of bias was due to idiosyncratic items and would not generalize to a wider-range of scenarios.

A different possibility is suggested by the fact that Tversky and Kahneman (1980) asked their participants to choose which of two probabilities is higher while we asked people to estimate probabilities for individual questions. To assess whether this procedural difference matters, we attempted to replicate one of Tversky and Kahneman's examples with our procedure. We asked 20 people to estimate the likelihood that a daughter has blue eyes given that her mother does, and another 20 to estimate the likelihood that a mother has blue eyes given that her daughter does. We found no evidence for a bias in the predictive direction. D was actually rated higher than P ($M_D = 49.9$; $M_C = 42.5$), but this difference was not significant, $t(38)=1.02$, $p = .31$.[4] Tversky and Kahneman's finding did not generalize to a direct probability judgment task. This prompts the question of which task is more akin to real-world judgment and decision making. Our feeling is that while people may sometimes compare which of two probabilities is larger, evaluating the likelihood of a unitary event is probably more common and more natural.

Fn4

Models of Inductive Reasoning

We now consider whether existing models of inductive reasoning can account for our results. One class of models is based on causal Bayes nets. Such models incorporate causal structure and make normative predictions based on that structure and various parametric assumptions. One example was proposed by Shafto et al. (2008) who described a food web model to make predictions about transmitted predicates and a taxonomic model to make predictions about genetic properties. The model predicts an asymmetry favoring the predictive direction but only for transmitted predicates. This asymmetry always holds in the networks that they tested because the background transmission rate, analogous to our strength of alternatives, was held constant across all nodes in the network. In our materials, we manipulated the strength of alternatives, thereby reversing the asymmetry. A generalization of Shafto et al.'s model that allowed for different background rates would be consistent with our normative formulation, though it would not predict the neglect of alternative causes in predictive reasoning.

A similar analysis may explain the causal asymmetry reported by Medin et al. (2003). It suggests that psychological principles like ease of reasoning are not necessary to explain the phenomenon because predictive judgments should usually be stronger than diagnostic ones. On the basis of 10,000 samples taken from the joint uniform distribution over all three parameters, we found that P is greater than D in 65% of the parameter space. This implies that predictive judgments should tend to be higher than diagnostic ones and suggests that the asymmetry reported by Medin et al. may be a result of differences in the evidential value of the premises in the two directions of reasoning. Even though our results show that on balance people underestimate predictive judgments relative to diagnostic judgments, the informational asymmetry in Medin et al.'s materials may have been sufficient to yield higher judgments in the predictive direction.

What about other models of inductive reasoning that are not based on causal structure? Can they account for our results? In similarity-based models such as the similarity-coverage model (Osherson et al., 1990), the feature-based model (Sloman, 1993), and more recent Bayesian models (Heit, 1998; Sanjana & Tenenbaum, 2003; Tenenbaum & Griffiths, 2001), inductive strength is proposed to be a function of the similarity between the categories in the argument and no differential predictions are based on predicate differences. These models do sometimes predict asymmetries in arguments, but these asymmetries are driven by the typicality or distinctiveness of the categories and not by the causal structures suggested by predicates. The manipulation of strength of alternatives in our experiments kept categories constant while varying the predicate, and we found that judgments are usually higher with strong alternatives but diagnostic judgments are usually higher with weak alternatives. Similarity-based models could only account for these results if similarity asymmetries between categories vary as a function of the predicate such that this pattern emerges. We can think of no theory of similarity that would predict such a pattern. Of course, if a measure of similarity were based on causal structure, such a pattern might be predicted, but we see this as consistent with our claim that causal structure underlies probability judgment. This argument also applies to asymmetric geometric models of similarity (Krumhansl, 1978).

Smith, Shafir and Osherson (1993) proposed the Gap model to account for arguments about nonblank predicates that violate the predictions of similarity-based models. For instance, the argument stating, "Poodles can bite through wire; therefore, German shepherds can bite through wire" was rated as a stronger argument than the one stating, "Collies can bite through wire; therefore, German shepherds can bite through wire" despite the fact that poodles are less similar to German shepherds than are collies. The idea behind the model is that a more surprising or implausible premise increases the conditional probability of the conclusion because it leads to greater belief revision. The fact that poodles can bite

[4] Methods were as follows: Participants were approached on the Brown University campus and took part voluntarily. They were asked either the predictive or diagnostic question verbally, and the experimenter wrote down their responses. Responses were analyzed with an independent samples t test.

through wire is more surprising than the fact that collies can, and this leads to more change in belief about German shepherds. Blok, Medin, and Osherson (2007) further developed this idea with the SimProb model, according to which the conditional probability of a conclusion for a one-premise argument is as follows:

$$P(Conclusion \mid Premise) = P(Conclusion)^{\left[\frac{1-SIM(\ premise,conclusion)}{1+SIM(\ premise,conclusion)}\right]^{1-P(\ premise)}} \quad (5)$$

where *SIM(premise, conclusion)* is the similarity between the premise and conclusion categories, which varies between 0 and 1 and is maximal at 1. The intuition behind the equation is that conditional probability is a joint function of the similarity of the premise and conclusion categories, and the plausibility of the premise, which is represented by $1 - P(Premise)$. We performed an in-depth analysis of SimProb's fit to our data, which revealed that SimProb is unable to account for our results (see Appendix B).

Neglect of Alternative Causes

The neglect of alternative causes that we found adds to a substantial literature showing that people often make errors of myopia, focusing unduly on a hypothesis currently under consideration while ignoring relevant alternatives. For instance, using an inductive inference task with uncertain premises, Hadjichristidis, Sloman, and Over (2009) found that people update their belief that a conclusion category has some property in a way that overweights the possibility that the premise is true relative to the possibility that it is false. The effect is reminiscent of pseudodiagnosticity (Doherty, Chadwick, Garavan, Barr, & Mynatt, 1996; Doherty, Mynatt, Tweeney, & Schiavo, 1979). To test a hypothesis, people tend to choose conditional probabilities involving hypotheses that they believe to be true rather than conditional probabilities that would actually support a comparison with alternative hypotheses. Using a different inductive inference task in which participants made predictions about events in stories, Ross and Murphy (1996) found that participants considered only the most likely character to be involved in the event, neglecting other characters. Reviewing the literature on how people test hypotheses and prior work on confirmation bias (e.g., Lord, Ross, & Lepper, 1979), Klayman and Ha (1987, p. 212) proposed that people apply a "positive test strategy" according to which they "test a hypothesis by examining instances in which the property or event is expected to occur (to see if it does occur), or by examining instances in which it is known to have occurred (to see if the hypothesized conditions prevail)." Evans, Over, and Handley (2003) introduced the *singularity principle* to describe this propensity to neglect alternative hypotheses. The principle implies that people tend to focus on only a single source when making an inference. One thing that these studies have in common is that participants were not asked to make an explicit judgment of diagnostic likelihood. Participants could by default have adopted a predictive mindset. For instance, in hypothesis testing, it may be most natural to think forward from a potential test to the likely outcome of that test, as opposed to thinking diagnostically from the possible outcomes of the test to the likely causes. Our work suggests that such a diagnostic mindset might lead to broader thinking.

In experiments in which participants are explicitly asked for diagnostic judgments, performance is in line with our findings of consideration of alternatives and consistency between beliefs and judgments. For instance, Dougherty, Gettys, and Thomas (1997) gave people vignettes describing a set of events and an outcome and asked for diagnostic judgments of the likelihood of some cause. In one example, participants read a story describing a fireman's death and judged the probability that it was caused by smoke inhalation. People who thought of alternative causes for death gave lower diagnostic judgments than those who did not. People tended not to think of many alternative causes, but this may be because the vignettes made alternative causes seem very unlikely. Another example comes from Waldmann (2000) who explored diagnostic reasoning in the context of a causal learning paradigm. Participants who learned about two possible diseases that could cause a symptom gave lower diagnostic judgments than those who learned about only a single cause. A challenge comes from Fischhoff, Slovic, and Lichtenstein's (1978) well-known troubleshooting study. They found that participants (and even experts) did not think of a fully comprehensive set of alternative causes for an engine failure, suggesting that their diagnostic judgments might not be perfectly calibrated with the true probability distribution. Participants did, however, think of a variety of alternatives, many of the most important ones, in line with our findings.

Previous studies in which forward to backward inference has been compared have also tended to find the same pattern. In a study of conditional reasoning, Cummins (1995; also see Cummins, Lubart, Alksnin, & Rist, 1991) found that participants gave higher acceptability ratings to affirming the consequent (AC) arguments about causal scenarios with few alternative causes. For instance, an argument like, "If the trigger was pulled, then the gun fired. The gun fired. Therefore, the trigger was pulled" obtained high ratings relative to "If Mary jumped in the pool, then she got wet. Mary got wet. Therefore, Mary jumped in the pool." AC is a logical fallacy because on the assumption that "if . . . then" refers to a material conditional, the presence of the consequent does not imply the antecedent. Yet when interpreted causally, AC is similar to *D,* in that it requires reasoning from effect to cause. However, judgments of modus ponens, which are analogous to *P*, were insensitive to alternative strength. Fernbach, Darlow, and Sloman (2010) obtained similar findings in a probability judgment task when manipulating the presence of alternative causes directly.

Conclusions

Diagnostic judgments are inherently comparative in the sense that they are, in part, a measure of how likely the target cause was to have brought about the effect relative to other causes. In the most direct kind of diagnostic task, a judgment of the conditional probability of a cause given an effect, people make this comparison and give responses that closely approximate the normative calculation. In contrast, people neglect alternatives when generating predictive probabilities and hence underestimate the likelihood of effects, even though they take alternatives into account if reminded to do so.

In some ways, this is a paradoxical result; Diagnostic reasoning is more complex in that it requires considering all three factors—prior probability, causal power and alternatives—while predictive reasoning is a simpler function of two of them. This suggests that the stumbling block to good inductive reasoning is not the complexity of the required computations. People have the capacity to

make good judgments when they consider the right factors, but they fail to take into account all that they should.

Our conclusion is that here are two contributors to differences in probability judgments based on causal directionality: One is the normative considerations highlighted by the Bayes net analysis: causes and effects often provide asymmetric evidential value for one another. The other is the nonnormative considerations based on people's tendency to think about only the focal causal mechanism when making predictions and the attendant ease with which they make such judgments. There is no simple "causal asymmetry" bias in the sense of probability flowing from cause to effect like water flowing down an incline. Like water, probability can be made to flow uphill; it just does not happen naturally but takes work.

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. doi:10.1037/0022-3514.35.5.303

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136,* 569–576. doi:10.1037/0096-3445.136.4.569

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

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

Blok, S. V., Medin, D. L., & Osherson, D. N. (2007). Induction as conditional probability judgment. *Memory & Cognition, 35,* 1353–1364.

Bonini, N., Tentori, K., & Osherson, D. (2004). A different conjunction fallacy. *Mind and Language, 19,* 199–210. doi:10.1111/j.1468-0017.2004.00254.x

Brem, S. K., & Rips, L. J. (2000). Evidence and explanation in informal argument. *Cognitive Science, 24,* 573–604. doi:10.1207/s15516709cog2404_2

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74,* 271–280. doi:10.1037/h0027592

Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for Bayesian cognitive science.* Oxford, England: Oxford University Press.

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

Cummins, D. D. (1995). Naïve theories and causal deduction. *Memory & Cognition, 23,* 646–658.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition, 19,* 274–282.

Doherty, M. E., Chadwick, R., Garavan, H., Barr, D., & Mynatt, C. R. (1996). On people's understanding of the diagnostic implications of probabilistic data. *Memory & Cognition, 24,* 644–654.

Doherty, M. E., Mynatt, C. R., Tweeney, R. D., & Schiavo, M. D. (1979). On pseudodiagnosticity. *Acta Psychologica, 43,* 111–121. doi:10.1016/0001-6918(79)90017-9

Dougherty, M. R. P., Gettys, C. F., & Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes, 70,* 135–148. doi:10.1006/obhd.1997.2700

Dowe, P. (2000). *Physical causation.* Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511570650

Evans, J. St. B. T., Over, D. E., & Handley, S. J. (2003). A theory of hypothetical thinking. In D. Hardman & L. Maachi (Eds.), *The psychology of reasoning and decision making* (pp. 3–21). Chichester, England: Wiley.

Fernbach, P. M., & Darlow, A. (2010). Causal conditional reasoning and conditional likelihood. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the Cognitive Science Society* (pp. XX–XX). Austin, TX: Cognitive Science Society. AQ: 1

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

Fischhoff, B., Slovic, P., & Lichtenstein (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception, and Performance, 4,* 330–344. doi:10.1037/0096-1523.4.2.330

Fox, C. R., & Levav, J. (2004). Partition-edit-count: Naïve extensional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General, 133,* 626–642. doi:10.1037/0096-3445.133.4.626

Galinsky, A. D., & Moskowitz, G. B. (2000). Counterfactuals as behavioural primes: Priming the simulation of heuristics and consideration of alternatives. *Journal of Experimental Social Psychology, 36,* 384–409. doi:10.1006/jesp.1999.1409

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704. doi:10.1037/0033-295X.102.4.684

Gilovich, T., & Griffin, D. (2002). Heuristics and biases: Then and now. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.). *Heuristics and biases: The psychology of intuitive judgment* (pp. 230–249). Cambridge, England: Cambridge University Press.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology.* Cambridge, MA: Bradford.

Goodman, N. (1955). *Fact, fiction, and forecast.* Cambridge, MA: Harvard University Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 3–32. doi:10.1037/0033-295X.111.1.3

Gopnik, A., & Schulz, L. E. (2007). *Causal learning: Psychology, philosophy, and computation.* Oxford, England: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334–384. doi:10.1016/j.cogpsych.2005.05.004

Hadjichristidis, C., Sloman, S. A., & Over, D. E. (2009). *Categorical induction from uncertain premises: Jeffrey's (doesn't) rule*. Manuscript submitted for publication.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248–274). Oxford, England: Oxford University Press.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 411–422. doi:10.1037/0278-7393.20.2.411

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12,* 275–305. doi:10.1002/(SICI)1099-0771(199912)12:4<275::AID-BDM323>3.0.CO;2-M

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. (1999), Naïve probability: A mental model theory of extensional reasoning. *Psychological Review, 106,* 62–88. doi:10.1037/0033-295X.106.1.62

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York, NY: Cambridge University Press.

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

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116,* 20–58. doi:10.1037/a0014282

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228. doi:10.1037/0033-295X.94.2.211

Krumhansl, C. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review, 85,* 445–463. doi:10.1037/0033-295X.85.5.445

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

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37,* 2098–2109. doi:10.1037/0022-3514.37.11.2098

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin and Review, 10,* 517–532.

Osherson, D. M., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97,* 185–200. doi:10.1037/0033-295X.97.2.185

Pearl, J. (2000). *Causality.* Cambridge, England: Cambridge University Press.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1141–1159. doi:10.1037/0278-7393.29.6.1141

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

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science, 33,* 301–344. doi:10.1111/j.1551-6709.2009.01015.x

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14,* 665–681. doi:10.1016/S0022-5371(75)80055-7

Rips, L. J. (2001). Necessity and natural categories. *Psychological Bulletin, 127,* 827–852. doi:10.1037/0033-2909.127.6.827

Ross, B. H., & Murphy, G. L. (1996). Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 736–753. doi:10.1037/0278-7393.22.3.736

Sanjana, N. E., & Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. *Advances in Neural Information Processes System, 15,* 51–58.

Shafto, P., Kemp, C., Baraff Bonawitz, E., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition, 109,* 175–192. doi:10.1016/j.cognition.2008.07.006

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology, 25,* 231–280. doi:10.1006/cogp.1993.1006

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives.* New York, NY: Oxford University Press.

Sloman, S. A., Barbey, A. K., & Hotaling, J. (2009). A causal model theory of the meaning of "cause," "enable," and "prevent." *Cognitive Science, 33,* 21–50. doi:10.1111/j.1551-6709.2008.01002.x

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2010). A causal model of intentionality judgment. Manuscript submitted for publication.

Sloman, S. A., & Hagmayer, Y. (2006). The causal psycho-logic of choice. *Trends in Cognitive Sciences, 10,* 407–412. doi:10.1016/j.tics.2006.07.001

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

Sloman, S. A., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 573–582. doi:10.1037/0278-7393.30.3.573

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition, 49,* 67–96. doi:10.1016/0010-0277(93)90036-U

Sobel, D. M., Tenenbaum, J. B., & 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., & Scheines, R. (1993). *Causation, prediction and search.* New York, NY: Springer–Verlag.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–640. doi:10.1017/S0140525X01000061

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207–232. doi:10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974, September 27). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131. doi:10.1126/science.185.4157.1124

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

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky, (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge, England: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293–315. doi:10.1037/0033-295X.90.4.293

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* 547–567. doi:10.1037/0033-295X.101.4.547

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. doi:10.1037/0278-7393.26.1.53

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: A minimal rational model. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 453–484). Oxford, England: 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. doi:10.1037/0096-3445.121.2.222

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology, 56,* 161–169. doi:10.1037/0022-3514.56.2.161

(*Appendices follow*)

tapraid5/zfr-xge/zfr-xge/zfr00111/zfr2190d11z | xppws | S=1 | 12/7/10 | 6:12 | Art: 2009-0419

Appendix A

Categories and Predicates Used in Experiments 1 and 3

Cause category	Effect category	Strong alternatives predicate	Weak alternatives predicate
Experiment 1			
Mother	Newborn baby	Has dark skin	Is drug addicted
Parents in New York City	Only child	Speak(s) English as first language	Know(s) child's birthday present
Coach	High school football team	Is motivated	Knows a complicated play
Commuter train	Commuter	Is late	Passes through several stations
Machine for manufacturing lenses	Lens	Is defective	Has micrometer precision
Mayor of a major city	New policy	Is unpopular	Is fiscally conservative
Hard disk	Computer	Is broken	Cannot hold any more files
Wheels	Car	Fail(s) inspection	Move(s) fast
Television manufacturers	Electronics stores	Sold an above-average number of defective products in 2007	Introduced a TV based on a new standard in 2007
Oranges	Orange smoothie	Are/is sweet	Are/is sour
Apple slices used to make an apple pie	Apple pie	Are/is sweet	Have/has seeds
Music at a party	Party	Is loud	Is good for dancing
Company on the New York Stock Exchange	Senior manager at the company	Is doing well financially	Uses Blue Cross health insurance
Transfusion blood at African hospital	Transfusion patient	Has an infectious disease	Is anemic
Early spring day in New York City	An apartment in New York City	Is warm	Is sunny
Engine of a 2005 Honda accord	2005 Honda Accord	Is not functioning properly	Smells of burnt oil
Northern ash wood	Baseball bat made from the wood	Is dark in color	Is liable to split
Body of water	Stew made from fish that live in the body of water	Is salty	Is high in mercury
Oxygen tank	Scuba diver	Has insufficient oxygen	Has plenty of oxygen
Tap water	Ice cubes made from the tap water	Taste(s) bad	Contain(s) fluoride
Experiment 3			
Mother	Newborn baby	Has dark skin	A father with dark skin
Coach	High school football team	Is motivated	Accolades from family and friends
Hard disk	Computer	Is broken	Other parts of the computer being broken, like the power source or the motherboard
Television manufacturers	Electronics stores	Sold an above-average number of defective products in 2007	Defective products that come from other sources, like computer manufacturers
Apple slices used to make an apple pie	Apple pie	Are/is sweet	Sugar is added
Music at a party	Party	Is loud	Loud conversations or other sources of loud noise
Early spring day in New York City	An apartment in New York City	Is warm	A heater turned on
Northern ash wood	Baseball bat made from the wood	Is dark in color	Has dark paint or stain
Tap water	Ice cubes made from the tap water	Taste(s) bad	Frozen next to something that has a strong odor

(*Appendices continue*)

tapraid5/zfr-xge/zfr-xge/zfr00111/zfr2190d11z | xppws | S=1 | 12/7/10 | 6:12 | Art: 2009-0419

Appendix B

SimProb Predictions

Translating the SimProb equation to our materials, we can define equations for the SimProb predictions for P and D as follows:

$$P_{simprob} = P(Effect \mid Cause) = P(Effect)^{\left[\frac{1-SIM(Cause,Effect)}{1+SIM(Cause,Effect)}\right]^{1-P(Cause)}} = P_e^{\left[\frac{1-SIM(Cause,Effect)}{1+SIM(Cause,Effect)}\right]^{P_c}} \quad (6)$$

$$D_{simprob} = P(Cause \mid Effect) = P(Cause)^{\left[\frac{1-SIM(Effect,Cause)}{1+SIM(Effect,Cause)}\right]^{1-P(Effect)}} = P_c^{\left[\frac{1-SIM(Effect,Cause)}{1+SIM(Effect,Cause)}\right]^{1-P_e}} \quad (7)$$

In Experiment 2 we collected judgments of P_c and P_e so the only thing missing for fitting the SimProb equations was the similarity of the cause and effect categories in our arguments. We did not collect similarity judgments in the experiments, but the design of our studies introduced some constraints. We held categories constant across the strong/weak and predictive/diagnostic manipulations so the similarity judgments for categories should not vary as a function of question type or alternative strength. With the simplifying assumption that similarity is symmetric—that is, S*IM(effect, cause) = SIM(cause, effect)*)—we could apply the SimProb model to our data by introducing five similarity parameters, one for each category set used in the experiment.

We explored the parameter space by varying each of the parameters from .1 to .9 in increments of .1 and calculating the SimProb equations at each point using the mean values of P_c and P_e from Experiment 3 collapsed over participants. This resulted in model fits at 59,049 points in the space. For each point, we calculated the correlations between mean judgments of P and $P_{simprob}$ and between D and $D_{simprob}$. The mean correlation over all points for predictive judgments was .2 ($p = .5$), and for diagnostic judgments it was .003 ($p = .9$). The maximal values were .81 ($p = .005$) and .68 ($p = .03$), respectively. For comparison, the normative model was highly correlated with both predictive judgments, $r(8) = .86$, $p = .001$, and diagnostic judgments, $r(8) = .80$, $p = .005$. Despite the fact that the SimProb model had five free parameters versus zero for the normative model, its best fits were inferior to those achieved by the normative model, and on average, it was not significantly correlated with P or D.

In addition to the correlational analyses, we assessed the qualitative fit of SimProb to the main finding of Experiments 1 and 2, the interaction between judgment type and alternative strength. For four out of the five predictive questions in Experiment 3, P was judged higher for the strong than the weak predicate. For all five of the diagnostic questions, the weak alternatives predicate yielded higher judgments. In line with this finding, the normative model predicted that four of the five strong predicates would yield higher predictive judgments and that all five of the weak predicates would yield higher diagnostic judgments. Conversely, SimProb predicted on average that 4.1 of the strong predicates would yield higher diagnostic judgments. Moreover, for two categories, SimProb never predicted that the diagnostic judgment should be higher for the weak predicate. SimProb was better at matching the predictions of predictive arguments, where it predicted on average that 4.3 strong predicates would be higher. In other words, SimProb tended to predict that strong alternative items should yield higher predictive and diagnostic judgments while participants and the normative model generated the interaction.

We also tested SimProb by asking 12 people for the similarity parameters and using the mean of each parameter to fit the model. The mean values were .46 for *SIM(Mother, Newborn)*, .44 for *SIM(Coach, Team)*, .43 for *SIM(Apple Slices, Apple Pie)*, .40 for *SIM(Music at Party, Party)*, and .45 for *SIM(Engine, Honda Accord)*. The analysis yielded nonsignificant correlations to P, $r(8) = .37$, $p = .3$, and to D, $r(8) = .14$, $p = .7$, and both correlations were significantly lower than those of the normative model: for P, $p = .009$, and for D, $p = .006$. SimProb also predicted that all five of the strong alternatives items should yield higher predictive judgments and higher diagnostic judgments than the weak items, again inconsistent with the interaction.

In summary, SimProb failed to capture the qualitative result from Experiments 1 and 2, the interaction between question type and alternative strength. It also could not match the quantitative performance of the normative model, even with the advantage of five free parameters. It should be noted that SimProb is not aimed at modeling transmissions between premise and conclusion categories, and therefore it is not surprising that it cannot match the data. The results also do not imply that SimProb fails to capture reasoning about arguments of the type to which Blok et al. (2007) applied it. Our analyses do show however that reasoning about transmission predicates that draws on causal structure knowledge cannot be explained by premise plausibility.

Received December 17, 2009
Revision received October 20, 2010
Accepted October 21, 2010 ■