Psychological Review
2008, Vol. 115, No. 4, 955–984

0033-295X/08/$12.00 DOI: 10.1037/a0013256

Bayesian Generic Priors for Causal Learning

Hongjing Lu, Alan L. Yuille, Mimi Liljeholm, Patricia W. Cheng, and Keith J. Holyoak
University of California, Los Angeles

The article presents a Bayesian model of causal learning that incorporates *generic priors*—systematic assumptions about abstract properties of a system of cause–effect relations. The proposed generic priors for causal learning favor *sparse and strong* (SS) causes—causes that are few in number and high in their individual powers to produce or prevent effects. The *SS power* model couples these generic priors with a causal generating function based on the assumption that unobservable causal influences on an effect operate independently (P. W. Cheng, 1997). The authors tested this and other Bayesian models, as well as leading nonnormative models, by fitting multiple data sets in which several parameters were varied parametrically across multiple types of judgments. The SS power model accounted for data concerning judgments of both causal strength and causal structure (whether a causal link exists). The model explains why human judgments of causal structure (relative to a Bayesian model lacking these generic priors) are influenced more by causal power and the base rate of the effect and less by sample size. Broader implications of the Bayesian framework for human learning are discussed.

Keywords: causal learning, Bayesian models, strength judgments, structure judgments, generic priors

From a very young age, humans display a remarkable ability to acquire knowledge of the causal structure of the world (e.g., Bullock, Gelman, & Baillargeon, 1982), often learning cause–effect relations from just a handful of observations (e.g., Gopnik, Sobel, Schulz, & Glymour, 2001; Sobel & Kirkham, 2007). Causal knowledge is particularly valuable in guiding intelligent behavior, making it possible to make predictions, diagnose faults, plan interventions, and form explanations (see Buehner & Cheng, 2005). Rather than expending their limited cognitive resources in a vain effort to learn all possible covariations among events, humans appear to focus on the more tractable (but still daunting) task of learning which types of events produce (or prevent) other types of events. A basic question remains: How can people (and possibly other animals) acquire causal knowledge from limited observations?

Hongjing Lu, Mimi Liljeholm, Patricia W. Cheng, and Keith J. Holyoak, Department of Psychology, University of California, Los Angeles (UCLA); Alan L. Yuille, Department of Statistics, Psychology, and Computer Science, UCLA.

Preparation of this article was supported by UCLA (Hongjing Lu), the W. H. Keck Foundation (Alan L. Yuille), the National Institutes of Health (Grant MH64810 to Patricia W. Cheng), and the Office of Naval Research (Grant N000140810186 to Keith J. Holyoak). Experiment 3 was part of a doctoral thesis completed in the UCLA Department of Psychology by Mimi Liljeholm under the direction of Patricia W. Cheng. Preliminary reports of part of this research were presented at the 28th (Vancouver, British Columbia, Canada, July 2006) and 29th (Nashville, TN, August 2007) Annual Conferences of the Cognitive Science Society. We thank David Danks, David Lagnado, and Michael Waldmann for helpful comments on earlier drafts. Matlab code for the models presented here is available from the Web site of the UCLA Computational Vision and Learning Lab (http://cvl.psych.ucla.edu).

Correspondence concerning this article should be addressed to Hongjing Lu, Department of Psychology, University of California, Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095-1563. E-mail: hongjing@ucla.edu

The philosopher Charles Peirce (1931–1958) argued that human induction must be guided by "special aptitudes for guessing right" (Vol. 2, p. 476). One possible guide to guessing right is simplicity or parsimony. The admonition often called Occam's razor was succinctly stated by Isaac Newton (1729/1968) as the first of his "Rules of Reasoning in Philosophy": "We are to admit no more causes of natural things, than such as are both true and sufficient to explain their appearances" (p. 3). The concept of simplicity poses thorny philosophical problems (Sober, 2002, 2006), both in defining simplicity and in justifying its use as a guide to induction. Yet the concept has longstanding appeal and recently has been proposed as a unifying principle for cognitive science (Chater & Vitányi, 2003). Applying the simplicity principle to causal reasoning, Lombrozo (2007) showed that when assessing causal explanations of individual events, people prefer explanations based on fewer causes (also Lagnado, 1994).

As Lombrozo (2007) noted, individual events are explained by *causal tokens*, that is, specific events that are instances of causal regularities. Our central aim in the present article is to formalize and test the possible role of simplicity in the acquisition of causal regularities that hold between *types* of events (see Sosa & Tooley, 1993). Working within a Bayesian framework for causal learning (Griffiths & Tenenbaum, 2005), we model simplicity using *generic priors*—systematic assumptions that human learners hold about abstract properties of a system of cause–effect relations. These generic priors, which function to constrain causal learning, provide a middle ground between complete absence of domain knowledge and dependence on highly specific prior knowledge. Even when the domain is unfamiliar, if the environment provides data consistent with the learner's generic priors, then causal learning can be rapid—on the human scale.

To explain how causal knowledge may be acquired and used, recent theoretical work on causal learning has made extensive use of formalisms based on directed causal graphs, simple examples of which are shown in Figure 1. Within a causal graph, each directed arrow connects a node representing a cause to one of its effects,

where it is understood that the cause does not follow its effect (typically preceding it) and has the power to generate or prevent it. (For the set of assumptions defining causal graphs, see Gopnik et al., 2004; Pearl, 1988, 2000; Spirtes, Glymour, & Scheines, 2000.) Causal powers can be interpreted as being probabilistic (i.e., a cause may yield its effect with a probability < 1). Such graphical representations, often termed *causal models,* have been used extensively in work on causal reasoning and learning in philosophy (Reichenbach, 1956; Salmon, 1984), artificial intelligence (Pearl, 1988, 2000; Spirtes et al., 2000), and psychology (e.g., Cheng, 1997; Gopnik et al., 2004; Griffiths & Tenenbaum, 2005; Waldmann & Holyoak, 1992; Waldmann & Martignon, 1998; for a review, see Lagnado, Waldmann, Hagmayer, & Sloman, 2007).

Framed in terms of causal graphs, the essential question is, How are causal models derived from some combination of prior knowledge and new observations? Causal graphs lend themselves to the development of rational models based on Bayesian inference (Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001). The heart of Bayesian inference is Bayes' rule,

$$P(H|D)=\frac{P(D|H)P(H)}{P(D)}, \quad (1)$$

where H denotes a hypothesized state of the world and D denotes observed data. Conceptually, Bayes' rule provides a mathematical tool to model the inference step by calculating the posterior probability of a hypothesis, $P(H|D)$, from prior belief about the probability of the hypothesis, $P(H)$, coupled with the likelihood of the new data in view of the hypothesis, $P(D|H)$. Assuming the hypothesis is causal (i.e., it can be represented as a link in a directed graph of the sort shown in Figure 1), developing a Bayesian model further requires specification of relevant prior beliefs and of a generating function[1] linking causal hypotheses to data.

If causal inference has a rational basis, we would expect the resulting causal knowledge to enable the formulation of coherent answers to a variety of causal queries (Cheng, Novick, Liljeholm, & Ford, 2007). Two major types of queries about causal links can be distinguished. One major type is, What is the probability with which a cause produces (alternatively, prevents) an effect? (E.g., for Graph 1 in Figure 1, this probability is the weight w_1 on the link from C to effect E.) This type of judgment, concerning the weight on a causal link, has been termed a *causal strength judgment* (e.g., Buehner, Cheng, & Clifford, 2003). Within a Bayesian framework, strength judgments pose a problem of *parameter estimation.* Another major type of query is, How likely is it that a causal link exists between these two variables? That is, does the cause have a nonzero probability of producing (or preventing) the effect? This type of judgment, concerning the existence of a causal link rather than its specific strength value, has been termed a *structure judgment* (Griffiths & Tenenbaum, 2005). Within a Bayesian framework, structure judgments pose a problem of *model selection* (e.g., for the graphs in Figure 1, the reasoner could decide whether Graph 1, which includes a link for candidate cause C, is more or less likely than Graph 0, which does not; see Mackay, 2003).

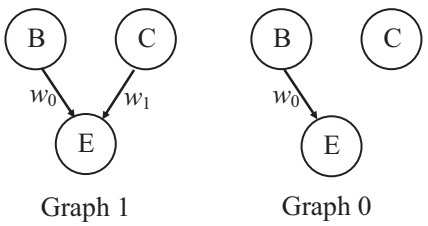

Figure 1. Graphs contrasting hypotheses that the candidate cause, C, causes the effect, E (Graph 1), or does not (Graph 0). B, C, and E denote the background cause, the candidate cause, and the effect, respectively. B, C, and E are binary variables that represent the absence and presence of the cause and the effect. Weights w_0 and w_1 indicate causal strength of the background cause (B) and the candidate cause (C), respectively.

Goals of the Article

In the present article, we extend prior formal work on causal inference (Cheng, 1997; Griffiths & Tenenbaum, 2005), developing and assessing a new Bayesian model of both strength and structure judgments termed the *SS power model*. The principal novelty of the model is its introduction of generic priors that implement a form of simplicity preference. These sparse and strong (SS) priors enable the model to account for the rapid acquisition of causal knowledge when the data match the priors. They also provide an explanation of certain subtle asymmetries between judgments of generative and preventive causes.

To formulate a full Bayesian model that can generate predictions, the hypothesized priors must be coupled with assumptions about the computation of likelihoods and about the relationship between different causal queries. Accordingly, we also examine the form that reasoners tacitly assume for the function by which multiple causes are combined to produce or prevent a common effect (e.g., how the influences of potential causes B and C in Graph 1 combine to determine E). For the special case of binary variables (the focus of the present article), some theorists have advocated a simple additive function that yields ΔP (see Equation 6 below) as an estimate of causal strength (Allan, 1980; Jenkins & Ward, 1965). Under certain conditions, the ΔP rule is equivalent to Rescorla and Wagner's (1972) associative learning model (which has been advanced as a model of causal inference; Shanks & Dickinson, 1987) when learning is at asymptote (see Danks, 2003). An alternative generating function, equivalent to a logical noisy-OR gate (see Equation 2), follows from the assumption of independent causal influence in the causal power PC theory (i.e., power theory of the probabilistic contrast model; Cheng, 1997; Novick & Cheng, 2004).[2] The SS power model adopts the same generating function as the power PC model; however, we also assess alternative models based on a linear function.

In addition, we examine the relationship between judgments of strength and of structure. Although these are theoretically distinct, it has been suggested that human reasoners often confuse the two, perhaps evaluating structure when queried about strength (Grif-

[1] The generating function is also commonly referred to as the *generating model* (a term we avoid because it is used in other ways throughout this article) and has also been termed the *parameterization* for combining multiple link strengths (Griffiths & Tenenbaum, 2005).

[2] Pearl (1988) suggested that the noisy-OR function could be used for causal graphs. His interpretation of this function, however, was not in terms of the influences of multiple causes operating independently. Cheng (1997) and Novick and Cheng (2004) were the first in cognitive science to provide the independent-influence interpretation for causal graphs based on the noisy-OR and its preventive analogue, noisy-AND-NOT.

fiths & Tenenbaum, 2005; Lagnado et al., 2007). Indeed, Griffiths and Tenenbaum (2005) modeled human strength judgments as Bayesian model selection, rather than as parameter estimation. However, recent empirical work (Liljeholm, 2007) has indicated that the two types of judgments are in fact psychologically distinguishable. We argue that the two types of judgments indeed differ, serving related but distinct computational goals.

We employ a research strategy that might be termed *computational cognitive psychophysics,* modeling large data sets in which multiple qualitative and quantitative parameters (causal direction, power, base rate of effect, sample size) are systematically manipulated across distinct but related judgments (strength and structure). A similar strategy is commonly employed in vision research (e.g., parametrically manipulating Gaussian noise across tasks such as detection and discrimination of objects; see Barlow & Tripathy, 1997; Lu & Yuille, 2006) and has also been used previously in studies of causal learning (e.g., Kao & Wasserman, 1993; Wasserman, Elek, Chatlosh, & Baker, 1993; Wasserman, Kao, Van Hamme, Katagiri, & Young, 1996). However, no previous study has encompassed judgments of both strength and structure. A rich body of data involving different causal queries enables tests of detailed quantitative models.

Theoretical Background

In this section, we review the two most direct antecedents of the SS power model: the power PC theory (Cheng, 1997) and the causal support model (Griffiths & Tenenbaum, 2005). We then introduce the generic priors for SS causes.

Power PC Theory

The SS power model incorporates the core claims of the power PC theory (Cheng, 1997), which reconciles the Humean view that causal relations are not directly observable with the Kantian view that people hold prior beliefs about unobserved powers of causes to produce (or prevent) their effects. We state the key psychological claims in relation to the causal graphs shown in Figure 1. Graph 1 represents the causal hypothesis that an effect E may be caused by a typically unobserved background cause, B, by an observed candidate cause C (being either present or absent on a trial), or both. A judgment of causal strength requires using data to infer the weights on each causal link, focusing on w_1, a random variable representing the strength of the candidate cause C. A judgment of causal structure requires assessing whether the data are more likely to have been produced by Graph 1 or by Graph 0, where the latter represents the possibility that C has no causal link to E (equivalently, that $w_1 = 0$).

The power PC theory postulates that people approach causal learning with four general prior beliefs:

1. B and C influence E independently,
2. B could produce E but not prevent it,
3. The causal powers of B and C (i.e., w_0 and w_1) are independent of the frequency of occurrences of B and C, and
4. E does not occur unless it is caused.

Assumptions 1 and 2 serve as default hypotheses for the reasoner, adopted unless evidence discredits them (in which case, alternative models apply; see Cheng, 2000; Novick & Cheng, 2004). Assumptions 3 and 4 are viewed as essential to causal inference (i.e., human causal inferences would be wrong or ineffectual without these assumptions). Assumption 4 is supported by research showing that adults (Kushnir, Gopnik, Schulz, & Danks, 2003), preschool children (Gelman & Kremer, 1991; Schulz & Sommerville, 2006), and even infants (Saxe, Tenenbaum, & Carey, 2005) interpret events as having causes, even when the state of the causes is unobservable.

For the special case of binary variables, these assumptions of the power PC theory imply a specific generating function for contingency data (Cheng, 1997; Glymour, 2001). Let $+/-$ indicate the value of a binary variable to be 1 versus 0. For the situation in which background cause B and candidate cause C are both potential generative causes, the probability of observing the effect E is given by a noisy-OR function,

$$P(e^+|b, c; w_0, w_1) = w_0 b + w_1 c - w_0 w_1 bc, \quad (2)$$

where b,c $\in \{0,1\}$ denotes the absence and the presence of the causes B and C. For simplicity, we follow Griffiths and Tenenbaum (2005) in treating the background cause B as if it is always present, so that $b = 1$.[3] Variables w_0 and w_1 are causal strengths of the background cause B and the candidate cause C, respectively. In the preventive case, B is again assumed to be potentially generative (Assumption 2), whereas C is potentially preventive. The resulting noisy-AND-NOT generating function for preventive causes is

$$P(e^+|b, c; w_0, w_1) = w_0 b - w_0 w_1 bc. \quad (3)$$

For convenience, we refer to Equations 2–3 together as the *power generating function.*

Using the power generating function, Cheng (1997) derived quantitative predictions for judgments of causal strength. Causal power, q, is defined as a maximum likelihood (ML) point estimate of w_1, the causal strength of the candidate cause. When the above assumptions of the power PC theory are satisfied and, in addition, causes occur independently of each other, the predicted value of causal power for a generative cause is

$$q_G = \frac{\Delta P}{1 - P(e^+|c^-)}, \quad (4)$$

and the predicted value of power for a preventive cause is

$$q_P = \frac{-\Delta P}{P(e^+|c^-)}, \quad (5)$$

[3] More precisely, B represents an amalgam of enabling conditions (Cheng & Novick, 1991; Mackie, 1974) and of observed and unobserved background causes, the latter occurring with unknown frequencies that may vary from situation to situation (Cheng, 1997). With respect to the effect of headache, for example, B might include constant enabling conditions such as being alive and capable of sensation, together with intermittent causes of headache such as noise or hot weather. In the present article, we focus on situations in which participants believe that enabling conditions are constantly present and the frequency of occurrence of background causes remains constant. In this case, for simplicity, we adopt the convention of referring to B as if it were a single constant background cause.

where ΔP is simply the difference between the probability of the effect in the presence versus absence of the candidate cause, that is,

$$\Delta P = P(e^+|c^+) - P(e^+|c^-). \quad (6)$$

The term $P(e^+|c^-)$ in the denominator of Equations 4–5 is often termed the *base rate of the effect,* as it gives the prevalence of the effect in the absence of the candidate cause. The base rate determines the point estimate of w_0 in Graph 1 (see Figure 1). Note that the causal power for generative causes (Equations 2 and 4) versus preventive causes (Equations 3 and 5) is inherently asymmetrical with respect to the base rate of the effect (see Cheng et al., 2007). We refer to a base rate of 0 (generative case) or 1 (preventive case) as *optimal base rates* because these are the values that maximize the number of cases that, for any data set of fixed size, could potentially reveal a causal effect of candidate *C*.

Causal Support Model

Griffiths and Tenenbaum (2005; Tenenbaum & Griffiths, 2001) developed a Bayesian causal support model to account for judgments as to whether a set of observations (*D*) was generated by a causal graphical structure in which a link exists between candidate cause *C* and effect *E* (Graph 1) or by a causal structure in which no link exists between *C* and *E* (Graph 0). The decision variable is obtained from the posterior probability ratio of Graphs 1 and 0 by applying Bayes' rule:

$$\log\frac{P(Graph1|D)}{P(Graph0|D)} = \log\frac{P(D|Graph1)}{P(D|Graph0)} + \log\frac{P(Graph1)}{P(Graph0)}. \quad (7)$$

Griffiths and Tenenbaum (2005) defined "causal support" as the first term on the right of Equation 7 (log likelihood ratio),

$$\text{support} = \log\frac{P(D|Graph1)}{P(D|Graph0)}, \quad (8)$$

because the second term in Equation 7, the log prior odds, is assumed to be 0. Support (also termed the *Bayes factor*; Mackay, 2003) gives a measure of the evidence that data *D* provide in favor of Graph 1 over Graph 0. Note that causal support is interpreted as a continuous measure of confidence in the existence of a causal link, rather than as a binary decision between Graph 0 and Graph 1.

The likelihoods on graphs are computed by averaging over the unknown parameter values, causal strengths w_0 and w_1, which lie in the range (0, 1) and are associated with causes *B* and *C*, respectively,

$$P(D|Graph1) = \int_0^1\int_0^1 P(D|w_0,w_1,Graph1)P(w_0,w_1|Graph1)dw_0dw_1,$$

$$P(D|Graph0) = \int_0^1 P(D|w_0,Graph0)P(w_0|Graph0)dw_0, \quad (9)$$

where $P(D|w_0, w_1, Graph1)$ and $P(D|w_0, Graph0)$ are the likelihoods of the observed data given specified causal strengths and structures and $P(w_0, w_1|Graph1)$ and $P(w_0|Graph0)$ are prior probability distributions that model the learner's beliefs about the distributions of causal strengths given a specific causal structure (assumed to be uniform, reflecting complete ignorance about the parameter values). Griffiths and Tenenbaum (2005) based their support model on the power generating function[4] (Equations 2–3); thus (for binary variables), their model constitutes a Bayesian extension of the power PC theory. Griffiths and Tenenbaum noted that, "speaking loosely, causal support is the Bayesian hypothesis test for which causal power is an effect size measure: it evaluates whether causal power is significantly different from zero" (Griffiths & Tenenbaum, 2005, p. 359).

To evaluate the support model, Griffiths and Tenenbaum (2005) reported three experiments designed to elicit structure judgments with binary variables. However, sample size was not systematically manipulated, and none of the experiments included preventive causes. More recently, Liljeholm (2007) performed several experiments that revealed ordinal violations of the support model as an account of human judgments. Relative to the support model, human reasoners appeared to place greater emphasis on causal power and the base rate of the effect and less emphasis on sample size. In addition, some contingency conditions yielded specific differences due to causal direction (generative vs. preventive), which are not captured by the support model. The present article reports additional data assessing human structure judgments.

Generic Priors for Sparse and Strong Causes

When learners have no obvious reason to have specific priors about weights (e.g., the power of a novel medicine to prevent headaches), one might suppose that the priors are simply uniform, as assumed in the causal support model (Griffiths & Tenenbaum, 2005). It is possible, however, that even when the inputs are entirely novel, learners may be guided by generic priors—systematic assumptions about the abstract quantitative properties of a system. In the case of motion perception, for example, human judgments of velocity are guided by the prior that motion tends to be slow and smooth. This generic prior explains a wide range of visual illusions and motion perception phenomena (Lu & Yuille, 2006; Weiss, Simoncelli, & Adelson, 2002; Yuille & Grzywacz, 1988).

One plausible prior assumption about the general nature of causal relations in the world, causal simplicity (Chater & Vitányi, 2003), potentially manifests itself in multiple ways. These include a preference (ceteris paribus) for fewer causes (Lombrozo, 2007) and for causes that minimize complex interactions (Novick & Cheng, 2004). Our present aim is not to provide a full account of causal simplicity but rather to focus on key aspects of simplicity that appear to guide elemental causal induction. The basic claim is that people prefer causal models that minimize the number of causes in a given direction (generative or preventive) while maximizing the strength of each individual cause that is in fact potent (i.e., of nonzero strength). We incorporate this preference, defined

[4] Although Griffiths and Tenenbaum (2005) also discussed a version of the support model based on a linear generating function, all their reported model fits were based on the power generating function.

as SS priors, in the SS power model to reflect a preference for simple causal models.[5]

SS priors are defined over causal strengths (rather than over causal structures). Specifically, we formulate SS priors as distributions of causal strengths given a potential causal structure (i.e., weights w_0 and w_1 for *B* and *C,* respectively). For the generative case, the background *B* and candidate *C* are both potentially generative and hence implicitly compete as alternative causes. Accordingly, we set priors favoring SS generative causes with the prior distribution peaks for (w_0, w_1) in Graph 1 at (0, 1) (*C* is the sole strong cause) and (1, 0) (*B* is; see Figure 2A). We specify the priors using a mixture of distributions with exponential functions,

$$P(w_0,w_1|gen,Graph1)\propto(e^{-\alpha w_0-\alpha(1-w_1)}+e^{-\alpha(1-w_0)-\alpha w_1}), \quad (10)$$

where α is a parameter controlling how strongly SS causes are preferred. When α = 0, the prior follows a uniform distribution, indicating no preference for any values of causal strength. As shown in Figure 2A, for a generative prior distribution when α = 5, the two most favorable situations in the SS prior distribution are $w_0 = 1$, $w_1 = 0$ (indicating that only the background cause *B* generates the effect), and $w_0 = 0$, $w_1 = 1$ (indicating that only the candidate cause *C* generates the effect). The impact of the SS prior is that when two (or more) possible generative causes of *E* co-occur and one cause has a stronger statistical link to *E* than the other, the presence of the stronger cause will tend to reduce the judged strength of the weaker one (cf. Baker, Mercier, Vallée-Tourangeau, Frank, & Pan, 1993; Busemeyer, Myung, & McDaniel, 1993). Figure 3 shows an example of the posterior distribution of w_1 obtained given contingency data of $p(e^+|c^-) = 12/16$ and $p(e^+|c^+) = 16/16$, based on either SS or uniform priors (see Appendix A for derivation). If domain knowledge provides more specific priors, these will be integrated with the generic priors (by multiplication) in calculating the posterior distribution. (A simulation presented in Appendix C illustrates how specific priors can be integrated with generic priors.)

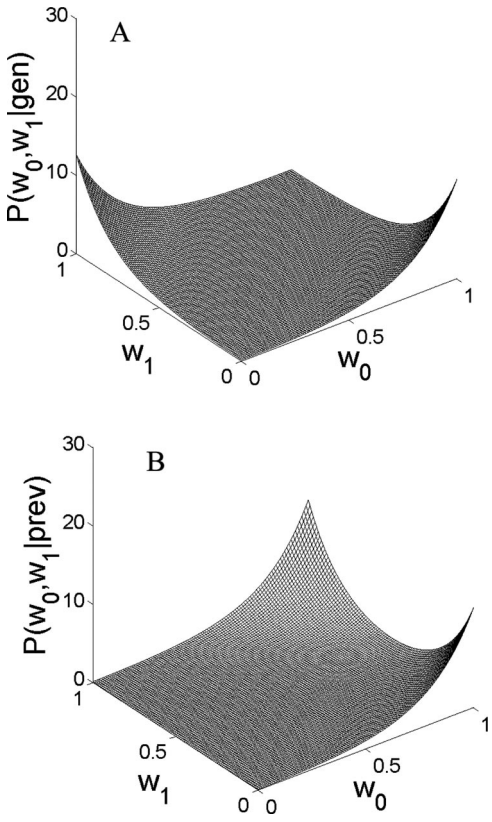

Figure 2. Prior distributions over w_0 and w_1 with sparse and strong priors. A: Generative case, α = 5 (peaks at [0, 1] and [1, 0]). B: Preventive case, α = 5 (peaks at [1, 1] and [1, 0]).

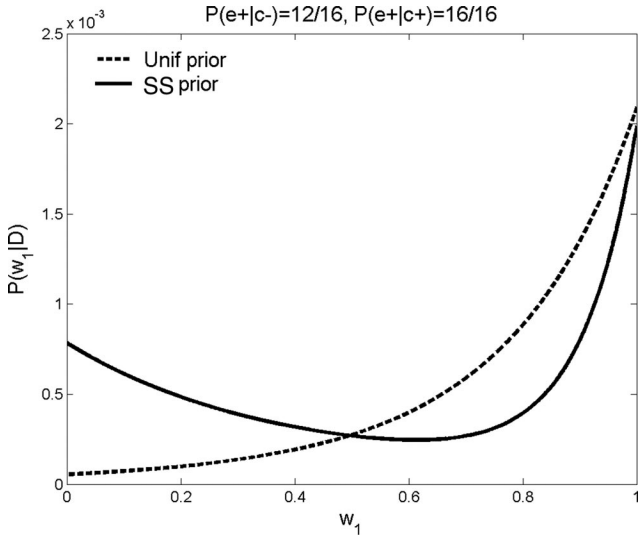

Figure 3. An example of a posterior distribution for w_1, given the contingency data $p(e^+|c^-) = 12/16$ and $p(e^+|c^+) = 16/16$. A: With uniform priors. B: With SS priors. SS = sparse and strong; Unif = uniform.

The SS prior will differ for the preventive case (see Figure 2B). Because the background cause, *B,* is assumed to be generative regardless of the existence of the preventive candidate cause *C, B* and *C* will not compete as alternative preventive causes. As *B* is the sole possible generative cause and the issue of prevention only arises when *E* is being generated, the peak weight of w_0 is assumed to be biased toward 1. The prior that *C* be a strong preventive cause then yields a distribution peak for (w_0, w_1) at (1, 1). If *C* (the only potential preventive cause) is not strong, the simplest alternative is that it is completely ineffective, yielding an additional peak at (1, 0). We again use an exponential formulation,

$$P(w_0,w_1|prev,Graph1)\propto(e^{-\alpha(1-w_0)-\alpha(1-w_1)}+e^{-\alpha(1-w_0)-\alpha w_1}), \quad (11)$$

[5] In preliminary reports of our model (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2006, 2007), we referred to these generic priors as favoring *necessary and sufficient causes,* a term that emphasizes the limiting ideal case of SS causes, namely, a single cause that is necessary and sufficient to produce or prevent its effect. However, the concepts of necessity and sufficiency have many usages both in discussions of causality (Sosa & Tooley, 1993) and in logic. The terms *sparse* and *strong* avoid several possible confusions; moreover, both terms clearly refer to continuous concepts (as sparseness and strength are each a matter of degree) and hence more accurately reflect the probabilistic nature of the proposed generic priors (see Equations 10–11).

where all parameters are defined as in Equation 10. If causal direction is unknown, then posteriors will be formed by averaging over the generative and preventive cases.

As shown in Figure 2B for a preventive prior distribution, the two most favorable situations in the preventive SS prior distribution are $w_0 = 1$, $w_1 = 0$ (indicating that only the background cause *B* generates the effect and that the candidate cause *C* is ineffective), and $w_0 = 1$, $w_1 = 1$ (indicating that the background cause *B* is a strong generator and the candidate cause *C* is a strong preventer). Note that the generative and preventive SS priors share the peak $w_0 = 1$, $w_1 = 0$, in which *C* is ineffective; however, the peak in which *C* is strong differs ($w_0 = 0$, $w_1 = 1$, for generative; $w_0 = 1$, $w_1 = 1$, for preventive). Because the latter two peaks are asymmetrical across causal direction, SS priors predict systematic asymmetries in human causal judgments. Two predicted asymmetries across causal direction are worth noting. First, for causes with relatively high strength, the preventive case will tend to be judged stronger than the generative case (for suitably matched contingencies) when the base rate is nonoptimal (i.e., close to 0 for preventive, close to 1 for generative). For example, a generative condition in which $P(e^+|c^-) = 0.75$, $P(e^+|c^+) = 1$ (power, the ML estimate of w_1, is 1 in this condition) is predicted to yield a lower strength estimate for *C* than the matched preventive condition in which $P(e^+|c^-) = 0.25$, $P(e^+|c^+) = 0$ (power is again 1). The reason is that the distance of the generative case (0.75, 1) from the (1, 0) peak (see Figure 2A) of the generative SS prior (which tends to reduce the estimated value of w_1) is less than the distance between the preventive case (0.25, 1) and the (1, 0) peak of the preventive prior (see Figure 2B).

Second, for causes with moderate strength, when the base rate is near the optimal value, judged strength is predicted to be greater for the generative case than for the matched preventive case. For example, a generative condition in which $P(e^+|c^-) = 0$, $P(e^+|c^+) = 0.25$ (power is 0.25), will tend to yield a higher strength estimate than the matched preventive condition in which $P(e^+|c^-) = 1$, $P(e^+|c^+) = 0.75$ (power is again 0.25). The reason is that the generative case (0, 0.25) is closer to the prior peak of $(w_0, w_1) = (0, 1)$ than to (1, 0), thereby biasing the estimate of w_1 toward 1. In contrast, the matched preventive case (1, 0.25) is closer to the prior peak of $(w_0, w_1) = (1, 0)$ than to (1, 1), biasing the estimate of w_1 toward 0. Thus, judgments of causal strength are predicted to be asymmetrical between preventive and corresponding generative conditions in certain regions of the contingency space.

The goals of strength and structure judgments differ in a manner that may influence their respective priors. A strength judgment focuses on the magnitude of w_1 within Graph 1, assuming (at least provisionally) that *C* may be a part of the causal topology; hence, neither *B* nor *C* is inherently favored over the other. In contrast, a structure judgment focuses on the question of whether or not candidate cause *C* should or should not be added to the topology of a causal graph in which *B* is already included. C thus has an inherent disadvantage, since sparseness clearly favors Graph 0 over Graph 1. It follows that the priors should support Graph 1 only if *C* is a strong cause, thereby justifying its addition to the set of accepted causes of *E*. We therefore assume that structure judgments reflect an additional preference that *C* (in Graph 1) be a strong cause of *E*. To construct the augmented SS+ priors for structure judgments, in both the generative and preventive cases we add a prior to favor $w_1 = 1$ (regardless of the strength of w_0),

$$P(w_0,w_1) \propto e^{-\beta(1-w_1)}, \tag{12}$$

where β is a parameter controlling the magnitude of this question-induced inductive bias. The higher the value of β, the stronger the preference that the strength of *C* be high. Figure 4 depicts the shape of a distribution for this prior when $\beta = 20$.

We then define the SS+ prior by simply multiplying the basic SS prior with the prior that C be strong,

$$P(w_0,w_1|gen,Graph1) = \frac{e^{-\beta(1-w_1)}(e^{-\alpha w_0-\alpha(1-w_1)}+e^{-\alpha(1-w_0)-\alpha w_1})}{Z_g(\alpha,\beta)} \tag{13}$$

and

$$P(w_0,w_1|prev,Graph1) = \frac{e^{-\beta(1-w_1)}(e^{-\alpha(1-w_0)-\alpha(1-w_1)} + e^{-\alpha(1-w_0)-\beta w_1})}{Z_p(\alpha,\beta)}. \tag{14}$$

$Z_g(\alpha, \beta)$ and $Z_p(\alpha, \beta)$ denote normalization terms for generative and preventive cases, respectively, which ensure that the sum of the prior probabilities over all possible values of w_0 and w_1 equals 1. Note that although the question-induced component of the SS+ prior does not differ across causal direction, the SS component does. Hence, the SS+ prior predicts systematic asymmetries between structure judgments for generative versus preventive causes, similar to the asymmetries across causal direction predicted for strength judgments.

In the area of vision, neural models that assume sparse coding (i.e., a sparse distribution of neural responses to natural images) have had considerable success (Graham & Field, 2007; Grimes & Rao, 2004; Olshausen & Field, 1996, 1997, 2004). Whereas pure sparse coding seeks to minimize all weights—equivalent to setting the weight peak for $[w_0, w_1]$ at [0, 0]), SS priors create a preference for finding the minimal set of weights that are relatively strong. The generic priors for SS causes can be viewed as a formalization

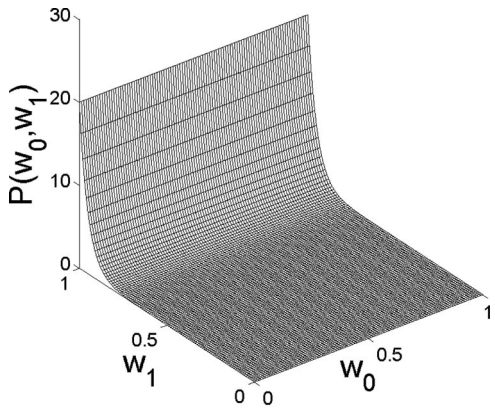

Figure 4

Figure 4. Prior distribution of question-induced inductive bias for structure judgments ($\beta = 20$).

of Newton's (1729/1968) first rule of reasoning in philosophy, mentioned earlier: "Admit no more causes of natural things, than such as are both true and sufficient to explain their appearances" (p. 3). That is, SS (and especially SS+) priors encourage acceptance of the smallest number of causes (sparse) that are in fact sufficient to explain the effect (strong). Besides the related proposal of Lombrozo (2007), the emphasis on sparseness of causes has precedent in Mackie's (1974) notion of a minimal sufficient condition for an effect, of which each component is an INUS condition (insufficient but nonredundant part of an unnecessary but sufficient condition). A minimal sufficient condition is a set of factors, minimal in number, that collectively suffice to produce a particular instance of an effect (but see Kim, 1993, for a critique of Mackie's, 1974, analysis of causation). The preference for strong causes is consistent with evidence that people selectively emphasize evidence that a contingency is sufficient (i.e., evidence that the two factors generally co-occur; Mandel & Lehman, 1998). Lien and Cheng (2000) proposed and provided evidence that a tacit goal of maximizing ΔP (i.e., maximizing the predictiveness of a single cause) guides human induction of categories and causal powers at multiple hierarchical levels.

The appropriate definition of simplicity and justification of its normative role in induction continue to be debated in philosophy. Sober (2002) argued that simplicity is a multifaceted concept that may have no global justification but rather multiple justifications in its various manifestations. For our present purposes, SS priors can be justified pragmatically by the basic assumption that cognitive capacity limits encourage satisficing strategies (Simon, 1955). Assuming that each additional cause carries a cost in processing load, one should rationally seek to minimize the number of assumed causes, with each accepted cause being maximally informative about the state of the effect. Informativeness of a cause can be defined in terms of the conditional entropy of the state of the effect given knowledge of the state of the cause. In the limit, a potential cause with 0 strength provides no information about the state of the effect beyond its base rate. In contrast, knowledge of a cause with strength of 1 yields a prediction that the effect is very likely to occur (generative cause) or not occur (preventive cause), thus reducing the conditional entropy associated with the effect.

It might be argued that SS priors are unrealistic as a description of the real world, which perhaps is better characterized by large numbers of weak causes. Indeed, Mill (1843) emphasized the problem for induction posed by the plurality of causes. On the other hand, Lewis (1979) claimed that events tend to have relatively few causes but many effects. In any case, a psychological theory of generic priors for causal learning can begin by sidestepping such ontological issues. Given the basic assumption of cognitive capacity limits, it is safe to assume that people will learn fewer causes more readily than more numerous causes and strong causes more readily than weak ones. It follows that whatever the state of causes in the world, the actual experience of a human learner will be biased toward acquiring SS causes, thereby reinforcing generic SS priors.

It should also be emphasized that SS priors do not preclude learning either multiple causes or veridical strengths. Like all effects of priors, the influence of SS priors is expected to be maximal early in learning, then to gradually disappear as learning approaches asymptote, because the influence of priors will eventually be swamped by the likelihoods based on the data when the sample becomes sufficiently large (Jaynes, 2003). Also, although priors for sparseness create competition among causes that co-occur (even if they are uncorrelated with each other; see Busemeyer et al., 1993), these priors need not impede learning for causes that do not co-occur. In Bayesian terms, the likelihood of any cause, however strong, producing its effect on occasions when that cause does not occur will be 0. Thus, learning that poison can be a cause of death will not interfere with learning that a gunshot can be a cause of death as long as shooting deaths are separate events from deaths by poisoning.

We now present a series of computational and empirical tests of the role of generic priors in causal learning. Table 1 provides a summary of the data sets we consider in evaluating alternative models.

Judgments of Causal Strength

To systematically compare alternative Bayesian models of judgments of causal strength, we implemented four models defined by the factorial combination of two alternative generating functions (power vs. linear) and two alternative priors (SS or uniform). The mathematical derivation of the models is presented in Appendix A. We refer to these alternatives as Model I (power, SS), Model II (power, uniform), Model III (linear, SS), and Model IV (linear, uniform). Model I corresponds to the SS power model when applied to estimate strength. Model II corresponds to the causal support model (Griffiths & Tenenbaum, 2005) when adapted to estimate causal strength (Danks, Griffiths, & Tenenbaum, 2003). Model IV corresponds to a Bayesian formulation of the ΔP rule (Jenkins & Ward, 1965) and the equivalent variant of the Rescorla-Wagner model (e.g., Shanks & Dickinson, 1987). Given the well-known empirical failures of the ΔP rule as an account of causal strength judgments (e.g., Buehner et al., 2003; Liljeholm & Cheng, 2007; Wu & Cheng, 1999), it would be surprising if Model IV were to provide the best account of the data considered in the present article. It is possible, however, that a model based on the linear generating function might be saved by augmenting it with generic SS priors. Accordingly, we also implemented Model III, which is identical to Model IV except with SS priors. We now compare the effectiveness of the four models as accounts of human judgments of causal strength.

Table 1
Summary of Data Sets for Judgments of Causal Strength and Structure Used to Evaluate Alternative Models

Judgments	Data set	Format
Strength judgments	Experiments 1, 2A, & 2B	Summary format
	Shanks (1995)	Sequential presentation
	Perales & Shanks (2007)	Sequential presentation (meta-analysis of 17 experiments from 10 studies)
Structure judgements	Experiments 3 & 4	Summary format
	Gopnik et al. (2001)	Sequential presentation (data from 4-year-old children)

Experiment 1: Varying Power, Base Rate, Sample Size, and Direction of Causation

Experiment 1 was designed to measure strength judgments across variations in causal power, base rate of the effect, sample size, and causal direction, providing data that could be used to assess alternative Bayesian models. To minimize memory issues and other factors extraneous to causal inference, Experiment 1 employed a procedure developed by Buehner et al. (2003) and extended by Liljeholm and Cheng (in press), in which individual trials are presented simultaneously in a single organized display (see Figure 5 for an example). Such presentations (which can be used to elicit either strength or structure judgments) provide a vivid display of individual cases, making salient the frequencies of the various types of cases while minimizing memory demands.

We also sought to elicit strength judgments using a procedure that minimizes ambiguity. Griffiths and Tenenbaum (2005) argued that people often confuse strength and structure judgments and proceeded to fit their causal support model to data in which participants were nominally asked to make strength (rather than structure) judgments (Buehner et al., 2003, Experiment 1; Lober & Shanks, 2000). Many studies have used numerical rating scales based on a variety of questions intended to assess causal strength (see Perales & Shanks, 2007). As pointed out by Buehner et al. (2003), such scales may be ambiguous, perhaps leading participants to assess the strength of the candidate cause specifically in the learning context.

An elicitation procedure for strength judgments that minimizes ambiguity is to ask participants to estimate the frequency with which the candidate cause would produce (or prevent) the effect in a new set of cases that do not already exhibit the effect (Buehner et al., 2003, Experiments 2–3). Measuring strength in a context in which no other cause is present should directly assess causal strength as it is theoretically defined, namely, the probability that the candidate cause would produce the effect. Strength judgments obtained using this procedure yield a pattern clearly more consistent with causal power than with causal support values, demonstrating that strength and structure judgments are empirically as well as theoretically separable (as acknowledged by Griffiths & Tenenbaum, 2005, pp. 374–375). We adapted a similar elicitation procedure for use in Experiment 1.

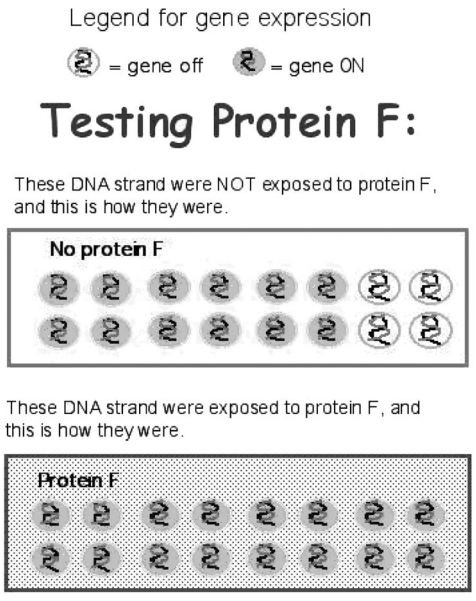

Figure 5. Example of an experimental display used in experiments with the DNA cover story, showing DNA strands that had not (top) or had (bottom) been exposed to a protein, resulting in a gene being on or off.

Method

Participants. Seventy-four University of California, Los Angeles (UCLA) undergraduates served in the study to obtain partial credit in an introductory psychology course. Participants were randomly assigned to conditions.

Materials, design, and procedure. The cover story concerned a bio-genetics company testing the influence of various proteins on the expression of a gene. Participants were told that, in each of several experiments, DNA strands extracted from hair samples would be exposed to a particular protein and that the expression of the gene would then be assessed. They were told that their job was to evaluate the influence of each protein on the expression of the gene. Each participant then saw a series of experiments, each of which showed two samples of DNA strands, depicted as vivid summaries (see Figure 5). One sample of DNA strands had not been exposed to a particular protein and depicted $P(e^+|c^-)$, while the other sample of DNA strands had been exposed to that protein and depicted $P(e^+|c^+)$.

The contingencies used in the experiment are shown in Figure 6. Contingency conditions were varied within subjects. The fractions in the two lines at the top and at bottom of Figure 6 indicate, respectively, the number of DNA strands that showed gene expression out of those not exposed to the protein (i.e., base rate of the effect) and the number that showed gene expression out of those that were exposed to the protein (where the protein is *C* and gene expression is *E*). The generative conditions (top two lines) and preventive conditions (bottom two lines) are identical except that the frequencies of gene expression and nonexpression are transposed. For example, the generative case 0/16, 4/16, where the base rate $P(e^+|c^-) = 0$, $P(e^+|c^+) = .25$, power = .25, and the sample size is 16, is matched to the symmetrical preventive case 16/16, 12/16, where $P(e^+|c^-) = 1$, $P(e^+|c^+) = .75$, power = .25, and the sample size is 16. In this and all experiments reported in this article, sample size is defined as the number of cases in which the cause was either present or absent (always an equal number).Thus, the sample size in Experiment 1 is coded as either 16 or 64.

Strength judgments were obtained from all participants. Causal direction was varied between participants, to ensure that they maintained a constant set to interpret problems as either generative or preventive. The causal query was modeled after that used by Buehner et al. (2003, Experiments 2–3), except that instead of a counterfactual wording ("imagine that there were a sample of 100. . . ."), we used a suppositional wording. Specifically, the causal query in the generative condition was

> Suppose that there is a sample of 100 DNA strands and that the gene is OFF in all those DNA strands. If these 100 strands were exposed to the protein, in how many of them would the gene be TURNED ON?"

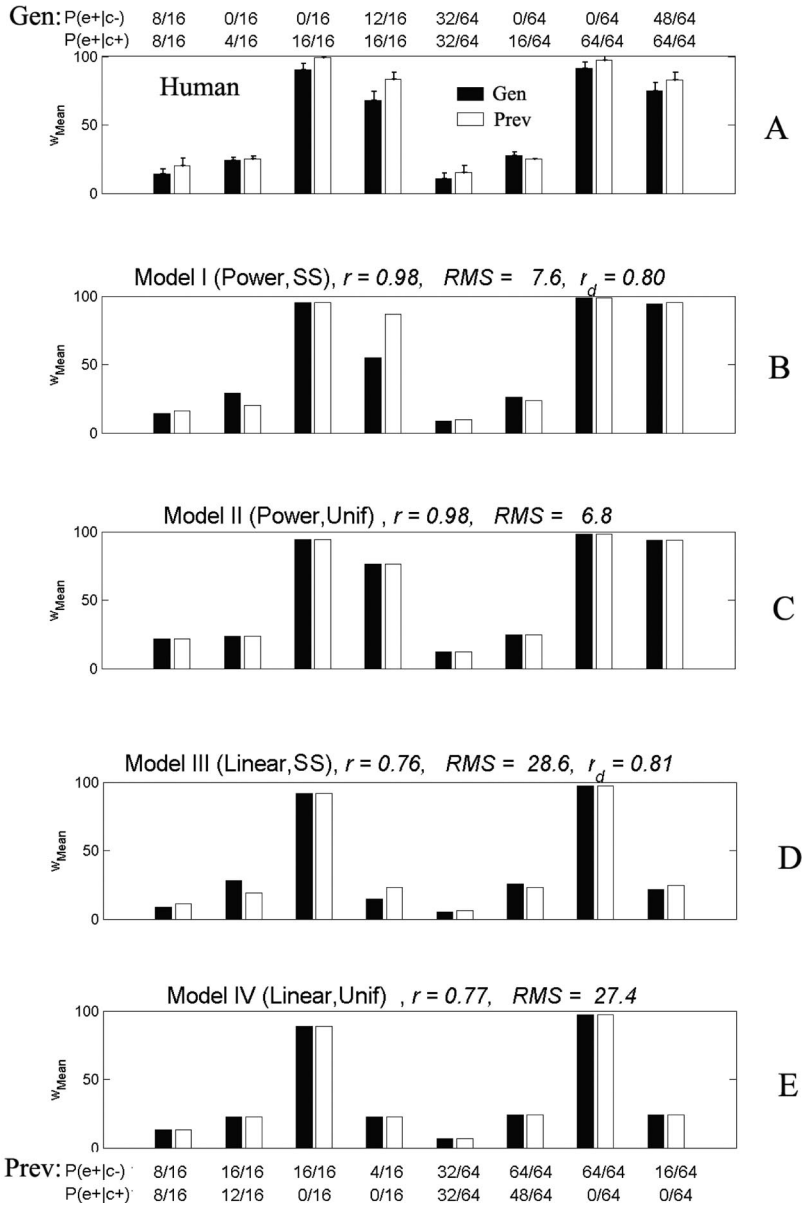

Figure 6. Judgments of causal strength (Experiment 1). A: Mean human strength judgments (error bars indicate one standard error). B: Predictions of the SS power model. C: Predictions of the model with power generating function and uniform priors. D: Predictions of the linear model with SS priors. E: Predictions of the linear model with uniform priors. Gen = generative; Prev = preventive; RMS = root-mean-square; SS = sparse and strong; Unif = uniform.

The preventive query was identical except that "OFF" was replaced by "ON" and "TURNED ON" by "TURNED OFF. Theoretically, the critical feature of this type of strength query is not the precise wording (counterfactual, suppositional, or actual) but rather that it measures the strength of the candidate cause in a context in which no other cause is producing the effect.

Results

Mean human strength judgments are shown in Figure 6A. An analysis of variance (ANOVA) was performed on these data. Strength judgments differed enormously across the four contingency conditions, $F(3, 216) = 205$, $p < .001$, reflecting the range of variations in predicted causal power (0–1). For matched contingencies, strength ratings were asymmetrical across the two causal directions, being reliably higher for preventive relative to generative causes, $F(1, 72) = 5.29$, $p = .024$. Although the data in Figure 6A suggest that the size of the preventive advantage tended to vary somewhat across conditions, no interactions were reliable. Judgments were virtually identical across the two sample sizes (16 vs. 64; $F < 1$). Similar findings concerning the impact of causal

direction on strength judgments have been reported by Liljeholm and Cheng (in press).

Comparison of Bayesian Models

Strength predictions were derived from the four Bayesian models, assuming a value of $\alpha = 5$ for SS priors. This value had originally been selected on the basis of an informal grid search using a different data set (see Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2007); by treating this parameter as fixed for all simulations reported in the present article, we avoided fitting a parameter to the various data sets (which would have favored the models with SS priors). For the models with uniform priors, the value of α is constrained to be 0. Predicted mean strength values can be derived from Bayesian models under the assumption that people estimate strength by implicitly sampling values drawn at random from the posterior probability distribution over w_1 (cf. Mamassian & Landy, 1998), a procedure leading to estimates that approach the mean of w_1 (see Appendix A, Equation A3). Accordingly, in our simulations, the mean of w_1 for each contingency was used to predict the corresponding mean strength rating. Following Buehner et al. (2003) and Liljeholm (2007), we assume that mean strength ratings on the 100-point scale approximate a ratio scale of causal strength.[6] Hence, a successful model must aim to account for the actual values obtained for human strength judgments, without any further data transformation. We therefore report model fits based not only on correlations but also on root-mean-square (RMS) deviations from the human data. In addition, the models with SS priors predict systematic differences as a function of causal direction. For Models I and III only, we therefore computed not only the overall correlation of model predictions with human data but also the correlation (r_d) between the observed and predicted difference between the mean strength judgments for matched generative and preventive contingencies. The predicted difference score is always 0 for Models II and IV, which assume uniform priors; hence, r_d is not computable. Because causal direction was manipulated as a between-subjects factor, differences involving causal direction tended to be more noisy than differences among the contingency conditions, which are based on within-subjects comparisons.

The human data based on the DNA cover story (see Figure 6A) were well fit overall by both of the models based on the power generating function, either Model I with SS priors (see Figure 6B) or Model II with uniform priors (see Figure 6C; $r = .98$ for each). The RMS was very low for both Models I and II, with a slight advantage for the model with uniform priors. However, the SS power model yielded a substantial positive correlation with the difference in strength ratings for matched generative and preventive contingencies ($r_d = .80$, $p < .02$), primarily attributable to conditions with nonoptimal base rates in which preventive strength was judged as exceeding generative strength for matched contingencies. The fit was nonetheless imperfect; for example, a predicted generative advantage for matched conditions with optimal base rate and low causal power—the generative condition where $P(e^+|c^-) = 0$, $P(e^+|c^+) = 0.25$, versus the matched preventive condition in which $P(e^+|c^-) = 1$, $P(e^+|c^+) = 0.75$—was not reliably obtained. Nonetheless, Model I was clearly more successful than Model II, as the model with uniform priors is completely unable to account for differences due to causal direction.

Models III and IV based on the linear generating function (see Figure 6D for SS priors, Figure 6E for uniform priors) yielded substantially poorer overall fits (rs = .77 and .76, respectively), roughly quadrupling the RMS relative to the models based on the power generating function. Model III, the linear model with SS priors, did yield a positive correlation with difference scores for generative versus preventive causes ($r_d = .81$, $p < .02$).

Note that Models I and II, based on the power generating function, succeed in capturing the higher mean strength ratings observed in the conditions in which $P(e^+|c^-) = 0$, $P(e^+|c^+) = 1.0$, than in those in which $P(e^+|c^-) = 0.75$, $P(e^+|c^+) = 1.0$ (especially at the smaller sample size). In both these paired conditions, power = 1, but mean ratings show an influence of the base rate of the effect and, hence, ΔP. Similar findings in the literature have previously been interpreted as evidence for use of the linear generating function (Lober & Shanks, 2000). Both Lober and Shanks (2000) and Buehner et al. (2003, Experiment 1) reported evidence of bimodality in the causal judgments of their participants, with at least a portion of them responding in accord with the ΔP rule.

Examining the judgments of individual participants in the present Experiment 1, we found that 8 out of 74 consistently gave responses that matched the ΔP rule. It is possible that this minority assumed a linear generating function that yields ΔP as its output. However, as Buehner et al. (2003) pointed out, it is also possible that some participants chose to answer the question of how much cause C increased the probability of E in the learning context, a query for which ΔP is the normative answer. In any case, dropping data from these participants scarcely altered the overall means across conditions ($r = .9996$ between means calculated with and without data from these participants), and in particular, mean ratings still varied with the base rate of the effect for conditions with equal causal power.

The relatively poor overall fits of Models III and IV imply that contrary to claims in the literature (Lober & Shanks, 2000), assuming the linear generating function does not provide an adequate account of these base rate effects. Instead, a more successful explanation of the influence of the base rate is provided by Models I and II, based on the power generating function coupled with Bayesian estimates of uncertainty. In these models, the base rate influences the posterior distribution of w_1 and, hence, its mean (see Appendix A, Equation A3). That is, when the base rate is relatively high (for generative causes), the mean of the posterior distribution shifts away from its peak (i.e., the point estimate of causal power) in the direction of ΔP. More generally, whenever the value of generative power differs from that of ΔP (equivalently, whenever the base rate is nonzero), the value of power (see Equations 4–5) is greater than the absolute value of ΔP (see Equation 6); hence, when uncertainty is introduced, the mean of the distribution of w_1 shifts away from power (the ML estimate of w_1) toward ΔP. By providing a quantitative account of uncertainty, the Bayesian mod-

[6] The assumption of a ratio scale is likely to break down for strength estimates near the extremes (0 or 100 on the scale) because of measurement issues. Whereas measurement errors near the middle of the scale can fall in either direction from the true mean and hence tend to cancel each other, measurement errors near the extremes can only overshoot (near 0) or undershoot (near 100), leading to systematic biases.

els thus clarify when and why mean strength judgments sometimes deviate from causal power.

The results of Experiment 1 indicate that the best overall Bayesian account of the pattern of human strength judgments is provided by Model I, the SS power model, which combines the power generating function with SS priors. The quantitative failure of the linear generating function (Models III and IV) confirms the negative conclusion that has been reached on the basis of ordinal comparisons (e.g., Buehner et al., 2003; Liljeholm & Cheng, 2007; Novick & Cheng, 2004). We thus can rule out the possibility that adopting the Bayesian framework might somehow salvage the linear generating function as a psychological model of human causal learning (see also Danks et al., 2003), regardless of whether the linear function is cast in terms of ΔP (Jenkins & Ward, 1965) or the Rescorla-Wagner model (Shanks & Dickinson, 1987).

Experiment 2: Does Expected Base Rate Differ Across Causal Directions?

The SS power model (Model I) implies that the prior on w_0 will differ as a function of the possible causal direction of candidate cause C. The two peaks of the priors for the preventive case both set $w_0 = 1$, whereas the peaks for the generative case differ, with w_0 set at either 1 or 0. The net effect is that the SS power model predicts that the expected base rate will be higher if the candidate cause is introduced as potentially preventive rather than potentially generative. For example, a scientist would be expected to test whether a medicine prevents some condition only if the condition is already known to occur with a substantial frequency (positive base rate) but might test whether a medicine generates some condition even if the condition is not yet known to occur (low or even zero base rate).

Experiment 2A

Experiment 2A was performed to assess whether simply specifying the possible direction of causal influence for C would indeed influence people's expectations about the base rate of the effect. We did not attempt to model the predicted difference quantitatively, as people may use different strategies to answer questions in the absence of any actual data. Some participants may not make use of generic priors in this task; instead, they may simply give low judgments because of the lack of data on which to base their estimates. Nonetheless, as long as at least some participants make use of generic priors to estimate the base rate in this task, a higher mean estimate should be obtained in the preventive than the generative condition.

Method

Eighty-one volunteers, ranging in age from 18 to 77 years old, participated in a Web-based experiment advertised on Craigslist.com. Forty-one people served in the generative condition and 40 in the preventive condition.

Causal direction was manipulated as a between-subjects variable, with participants randomly assigned to one of the two conditions by the computer. A computer display presented a cover story (generative condition) stating that,

> a pharmaceutical company is investigating whether a new allergy medicine might produce a medical condition called *malosis* as a side effect. Imagine you are a researcher in this company who will conduct a research study. You have been assigned 100 randomly-selected subjects who have agreed to take part in a clinical trial. Your job is to first determine whether or not each subject has malosis prior to administering the new medicine. Then you will administer the new medicine to all the subjects, and afterward again assess how many subjects have malosis.

The preventive cover story was identical except that in the first sentence, the word *prevent* replaced *produce*. No other information about cases of malosis was presented. The participant was then asked, "What is your best estimate of how many of the 100 subjects will have malosis BEFORE the medicine is administered? Give your best estimate, even if you are unsure." The computer recorded the numerical response (0–100).

Results and Discussion

The mean estimated base rate of malosis was 7.9 in the generative condition versus 22.4 in the preventive condition, $t(79) = 3.57$, $p < .001$. These results confirm that people's expectations about the base rate of an effect are sensitive to information about the potential causal direction of the candidate cause. As predicted by the SS power model, prior to seeing any data, the estimate of the base rate of the effect is greater when the candidate cause is introduced as potentially preventive, rather than generative.

Experiment 2B

Experiment 2B was performed to replicate the finding of Experiment 2A using a different cover story and participant population. It might be argued that with the drug-testing cover story of Experiment 2A, participants would expect the drug company to select patients who have the medical condition to test whether a drug prevents it. The mention of random sampling could have been interpreted as random sampling from among a population with malosis. The cover story used in Experiment 2B placed greater emphasis on random sampling and explicitly stated the population being sampled (which was identical for generative and preventive conditions).

Method

A total of 115 UCLA undergraduates enrolled in Introductory Psychology served in the study as part of course requirement. They were tested in a group that received a battery of pencil-and-paper tests during a 1-hr session. Experiment 2B took about 5 min to complete; the rest of the test battery consisted of unrelated questions. Sixty-one students served in the generative condition and 54 in the preventive condition.

Causal direction was manipulated as a between-subjects variable, with participants randomly assigned to one of the two conditions. The test booklet presented a cover story (generative condition) entitled "Testing a New Protein," which stated,

> Imagine you are a scientist investigating whether exposure to a new protein called *megdolin* might cause a gene called *CDR2* to be expressed in DNA strands taken from human hair samples. You have been given 100 DNA strands, selected by **random sampling** from

> hairs provided by all the soldiers stationed at a US army base. Your job is to first determine whether or not each DNA strand shows expression of the CDR2 gene prior to exposing it to megdolin. Then you will expose all DNA strands to megdolin, and afterward again assess how many samples show expression of the CDR2 gene.

The preventive cover story was identical except that in the first sentence, the word *prevent* replaced *cause*. Note that the population being sampled was clearly defined and identical for the two conditions.

The participant was then asked, "What is your best estimate of how many of the 100 DNA strands will show expression of the CDR2 gene BEFORE exposure to megdolin? Give your best estimate, even if you are unsure." The participant provided a numerical response (0–100).

Results and Discussion

The mean estimated base rate of gene expression was 25.2 in the generative condition versus 44.1 in the preventive condition, $t(113) = 3.47$, $p < .001$. The results of Experiment 2B thus replicated the basic finding of Experiment 2A using a different cover story, type of participants, and presentation procedure. In addition, the cover story in Experiment 2B emphasized that the sample cases were obtained by random sampling from a well-defined population that was identical in both conditions.

The results of Experiments 2A and 2B confirm that people's expectations about the base rate of an effect are sensitive to information about the potential causal direction of the candidate cause. It could be argued that the observed differences reflect a general expectation that scientists will test for a preventer when the effect is already known to be generated somehow (positive base rate), whereas they will test for a generator even if the effect is not yet known to be generated (lower base rate). Such an expectation would be consistent with the generic priors incorporated in the SS power model.

Learning Causal Strength From Sequential Presentation of Data

We have made a preliminary effort to extend Bayesian models of strength judgments to sequential learning—situations in which cases are presented one at a time so that causal parameters must be updated as data accrue. One of the apparent advantages of applying the Rescorla-Wagner model to causal learning is that it provides a natural account of many competitive effects observed in causal learning and also of the graded learning curve for acquiring knowledge of causal strength (Shanks & Dickinson, 1987). However, as we have seen, the Rescorla-Wagner model (as instantiated in Bayesian Model IV) incorrectly predicts that the asymptotic value of causal strength will approach ΔP, rather than causal power. Danks et al. (2003) were the first to develop models of sequential learning based on the power generating function (Cheng, 1997). Their models, to which the extensions we present here are closely related, avoid the fundamental failure of models based on the linear generating function.

As Danks et al. (2003) observed, any model of causal learning from summary data can be applied to sequential learning simply by keeping a running tally of the four cells of the contingency table, applying the model after accumulating n observations, and repeating as n increases. We take this tack here. By assuming perfect memory for the observations, we create models that constitute a type of Bayesian ideal observer (e.g., Barlow & Tripathy, 1997; Kittur, Holyoak, & Hummel, 2006; Lu & Yuille, 2006). It should be noted, however, that models of this type cannot account for either order effects or forgetting (although as a first step toward modeling forgetting, an exponential forgetting function for data could be introduced; see Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003.)

We developed sequential versions of Models I and II, both based on the power generating function but differing in their priors (SS priors for Model I, uniform for Model II). The models were applied to a set of data described by Shanks (1995), also simulated by Danks et al. (2003). In the human experiment, people saw a sequence of 40 trials of a computer-generated display. On each trial, a tank was either camouflaged or not and was either destroyed by a mine or not. After every five observations, the participants were asked to rate the causal strength of the camouflage on a scale ranging from −100 (strong preventer of destruction) to 100 (strong generative cause), with the midpoint of 0 indicating no causal relation. Different groups of participants received one of four contingencies. Two conditions had strong but imperfect contingencies, either in the generative direction with $P(e^+|c^-) = .25$, $P(e^+|c^+) = .75$, or in the preventive direction with $P(e^+|c^-) = .75$, $P(e^+|c^+) = .25$. These conditions were equated on absolute values of ΔP (.50) and causal power (.67). The remaining two conditions had zero contingencies, either with a high probability of the effect in which $P(e^+|c^-) = P(e^+|c^+) = .75$ or with a low probability of the effect in which $P(e^+|c^-) = P(e^+|c^+) = .25$. Importantly, and unlike the previous experiments we have modeled, participants were not informed about the possible causal direction but rather had to infer direction of causation from the observations.

Figure 7A displays the mean strength judgments reported by Shanks (1995). Three qualitative aspects of the results are of interest. (a) A negatively accelerated acquisition function was obtained for the nonzero-contingency conditions, as is typical of sequential learning (e.g., Shanks, 1987; Wasserman et al., 1993). (b) The two zero-contingency conditions differed across the early trials: The contingency with a higher effect probability—$P(e^+|c^-) = P(e^+|c^+) = 0.75$—initially received more positive strength ratings than did the contingency with a lower effect probability—$P(e^+|c^-) = P(e^+|c^+) = 0.25$—although both conditions eventually approached a mean rating of 0. Similar differences among zero-contingency conditions across which the probability of the effect was varied have been observed in other studies (e.g., Allan & Jenkins, 1983; Shanks, 1987; White, 2004). (c) There appears to have been a trend toward an asymmetry between the generative and preventive conditions (the two conditions with nonzero contingencies), with the generative condition showing a positive slope after Trial 10, whereas the slope for the preventive condition was relatively flat after 10 trials. It is unclear whether either condition reached asymptote after 40 trials. Assuming that the asymptote had not been reached, it appears the learning curve rose more quickly for the generative condition.

The sequential models that we developed are identical to Models I and II presented earlier, except for an extension required to handle the situation in which the causal direction is unknown. The

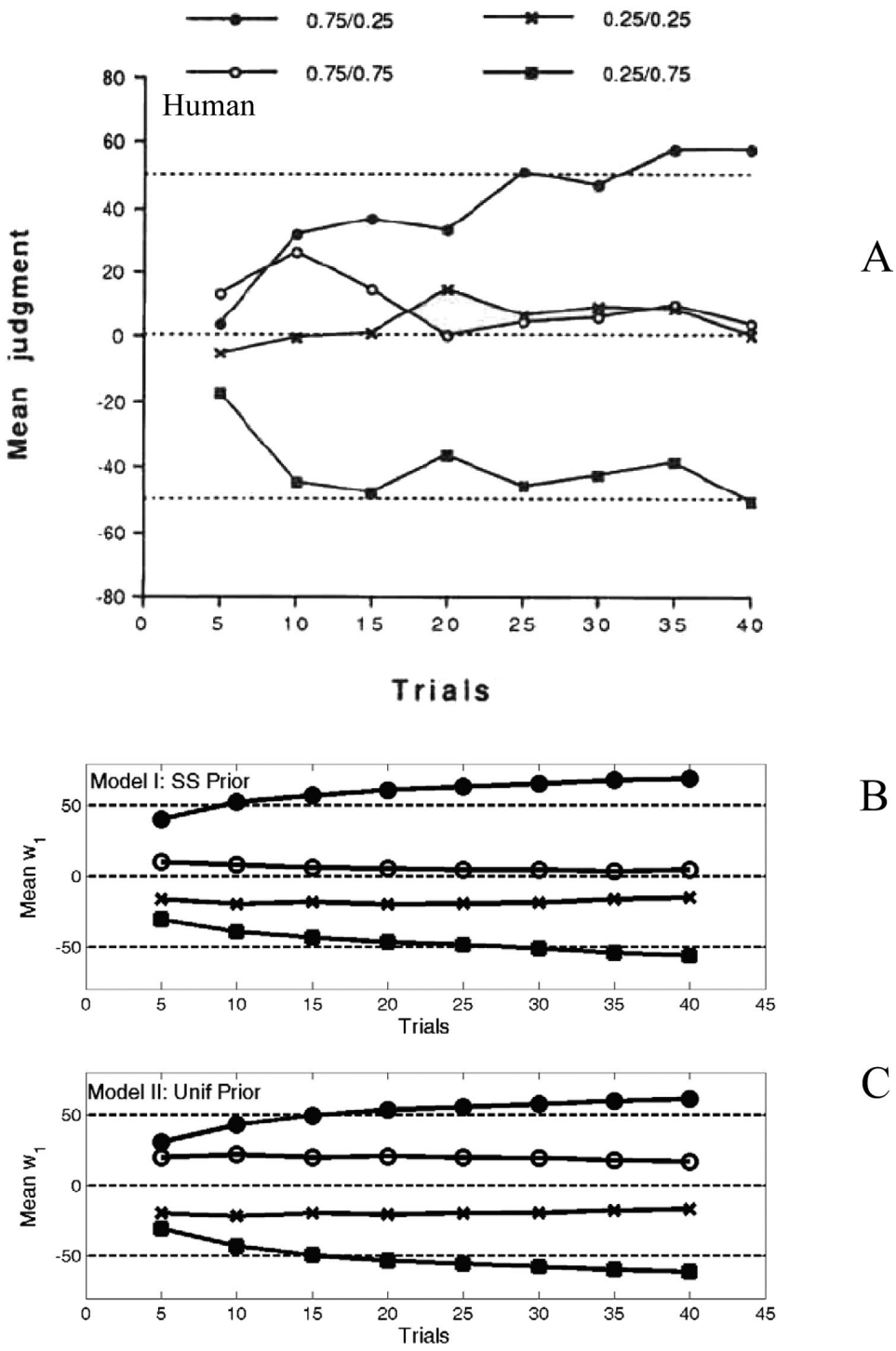

Figure 7. Mean causal strength judgments in a sequential learning paradigm. A: Human data (Shanks, 1995). B: Predictions based on SS priors (Model I). C: Predictions based on uniform priors (Model II). Both sets of simulation results are means of posterior distributions averaged over 200 runs for each condition. SS = sparse and strong; Unif = uniform. Figure 7A is from "Is Human Learning Rational?", by D. R. Shanks, 1995, *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *48*(A), p. 263. Copyright 1995 by the Experimental Psychology Society. Reprinted with permission.

models integrate over two possible causal directions,[7] reversing the sign for preventive causal strengths (thus matching the scoring used in the human experiment, in which preventive causes were coded as negative strength ratings). Predicted strength ratings are given by

$$\bar{w}_1 = \sum_{g \in G} \int_0^1 w_1 P(w_1|g,D) P(g|D) dw_1, \quad (15)$$

where $G = \{Graph1$ with generative cause, $Graph1$ with preventive cause$\}$. The sequential version of Model II (uniform priors) is virtually identical to one of the models constructed by Danks et al. (2003; see their Figure 2e), which also employed the power generating function with uniform priors over causal strengths. (In addition, Danks et al., 2003, reported simulation results assuming a nonuniform prior over w_0; see their Figure 2f.) The only difference is that our models (as in our previous simulations of human strength judgments) assume that in the experimental set-up, people base their strength judgments on Graph 1 (see Figure 1), provisionally accepting the possibility that the candidate C may be causal. In contrast, in the comparable models constructed by Danks et al., strength estimates were obtained by integrating over Graph 0 (in which $w_1 = 0$) as well as Graph 1. As a practical matter, integrating over Graph 0 as well as Graph 1 reduces the overall learning rate somewhat for nonzero contingencies but does not change the qualitative pattern of simulation results.

On each run of the simulations, data were generated stochastically according to the probability distributions for each contingency condition. The posterior distribution for w_1 was calculated after every five observations, and the mean of this distribution provided sequential estimates of causal strength. The simulation results (based on 200 runs for each condition) are shown in Figure 7B (Model I) and Figure 7C (Model II). Relative to the human data shown in Figure 7A, the Bayesian models learned more rapidly, as is to be expected given that they constitute ideal observers with perfect memory for observations. At a qualitative level, both models captured important aspects of the human data shown in Figure 7A: a negatively accelerating learning curve and greater positive strength (persisting even after 40 trials) for the zero-contingency condition in which the probability of the effect was relatively high. The latter phenomenon (sometimes viewed as a challenge for rational models of causal learning) is a natural consequence of the inherent asymmetry of the power generating function (noisy-OR for generative causes, noisy-AND-NOT for preventive causes) when applied to stochastic observations (Buehner et al., 2003). The net result will be an initial bias toward more positive strength estimates when the effect probability is relatively high.

It should be noted that whereas the human data show positive values on early trials for the zero-contingency condition with the lower base rate, both Bayesian models yielded slightly negative mean strength values. We have no clear explanation for this discrepancy, which may reflect a bias in participants' use of the rating scale.

As in the simulations of experiments based on summary data, Model II with uniform priors predicts that the generative and preventive conditions with nonzero contingencies will yield symmetrical causal strengths (see Figure 7C). In contrast, Model I with SS priors (see Figure 7B) yielded an asymmetry over the 40 trials, with a slightly faster learning rate for the generative than the matched preventive condition. Model I thus appears to have captured the trend toward an asymmetry across causal directions observed in the human data (see Figure 7A). As discussed earlier, a generative advantage is predicted by SS priors when the base rate is close to optimal and the value of w_1 is moderate, which is the case for the nonzero contingencies tested by Shanks (1995). If the model were run for a sufficiently large number of trials, the generative and preventive conditions would eventually converge at symmetrical strengths.

Although preliminary, the present formulations of sequential learning within Bayesian models, together with those reported by Danks et al. (2003), provide encouragement that the Bayesian approach has promise. We consider the possibility of creating more psychologically realistic sequential learning models in the General Discussion.

Meta-Analysis of Causal Strength Judgments

Perales and Shanks (2007) performed a meta-analysis of experiments that attempted to measure causal strength on the basis of binary contingency data, comparing predictions based on simple causal power, the ΔP rule, the Rescorla-Wagner model, causal support, and four additional nonnormative models. On the basis of fits to a total of 114 conditions taken from 17 experiments reported in 10 studies from multiple labs, Perales and Shanks concluded that the most successful model was a nonnormative model called the *evidence integration* (EI) *rule*. Hattori and Oaksford (2007) performed a similar meta-analysis comparing predictions derived from simple causal power, the ΔP rule, and 39 nonnormative models. Like Perales and Shanks, Hattori and Oaksford concluded that one of the nonnormative models was the most successful predictor of strength judgments; however, their favored model was not EI but instead a measure termed the *dual factor heuristic* (H).

Neither of these meta-analyses considered Bayesian models of strength judgments based on either the power or linear generating function. Following Griffiths and Tenenbaum (2005), Perales and Shanks (2007) treated causal support as an index of causal strength, reporting that it yielded the poorest fit of any of the models tested. This was true even though the meta-analysis included those experiments (Buehner et al., 2003, Experiment 1; Lober & Shanks, 2000) that Griffiths and Tenenbaum had fitted using the causal support model. (We note, however, that Perales and Shanks, 2007, failed to consider participants' knowledge of causal direction in calculating support values.) Of course, causal support does not provide a normative measure of causal strength. We consider whether strength and support judgments are empirically as well as conceptually distinct after we discuss models of causal structure in Experiment 3 below.

Fitting Bayesian Models to Meta-Analysis Data

We used the meta-analysis database provided as an appendix by Perales and Shanks (2007, pp. 594–596). To fit Bayesian models

[7] To model a situation in which the causal direction of C is unknown, the prior on w_0 in Graph 0 is set to the average of the marginal distribution obtained under the assumption that Graph 1 is (a) generative or (b) preventive.

to the meta-analysis data, it was necessary to take account of whether or not participants were informed about the possible direction of causation (generative or preventive). This information (not considered by Perales and Shanks, 2007) was obtained from the source articles. The models were then fit to each condition using the appropriate equations, exactly as described above for our own Experiment 1 (causal direction known) and for the data from Shanks (1995; causal direction unknown). As in these earlier simulations, the value of α was set to 5 for Model I (SS priors) and to 0 for Models II and IV (uniform priors); thus, the models were not parameter-fitted to the data sets included in Perales and Shanks's meta-analysis.

The correlations obtained for Models I and II, both based on the power generating function, were $r = .94$ (RMS = 14.64) and .96 (RMS = 10.95), respectively, across all 114 conditions included in the meta-analysis. These correlations and RMS values are comparable to or better than those obtained by Perales and Shanks (2007) using the EI rule, which has four free parameters ($r = .94$, RMS = 12.79). The fit of Model IV (linear generating function with uniform priors) was less adequate ($r = .91$, RMS = 20.02). The more subtle influence of generic priors on strength judgments that we observed in our own Experiment 1 was not detectable. The r_d measure reported for our Experiment 1 is not computable for the meta-analysis because generative and preventive conditions were not consistently equated for causal power and optimality of the base rate.

The good overall fits obtained for Models I and II are remarkable because some of the criteria that Perales and Shanks (2007) used in selecting experiments for their meta-analysis would seem to work against normative models. Their criteria led to inclusion of data that are problematic in various ways for evaluating causal inference. Only experiments using trial-by-trial presentation were included, thus introducing potential memory issues. Of greater concern, the only experiments included were those in which "standard causal questions were used (in which simply a general estimate of the relationship, and not a specification of the context in which the question applied, was required)" (Perales & Shanks, 2007, p. 584). In other words, all experiments in which the question clearly assessed causal strength (i.e., the probability that the candidate cause would produce the effect when acting alone) were excluded from the meta-analysis, whereas those experiments in which the question was open to various plausible alternative interpretations of what "a general estimate of the relationship" means, were included. For example, data from Experiment 1 in Buehner et al. (2003) were included because a vaguely worded query was used, whereas data from their Experiments 2–3, which used a much clearer query, were excluded. Data from Experiment 1 in the present article would of course be disqualified on the basis of the criteria used by Perales and Shanks.

Equally problematic, in many studies included in the meta-analysis, the instructions did not make it clear that alternative causes were held constant (e.g., that there was random assignment of cases to situations in which the candidate cause was present vs. absent). In contrast, studies that gave clear instructions regarding the independent occurrence of alternative causes were excluded (Buehner et al., 2003, Experiments 2–3). The Bayesian models we have considered are all derived from the assumption that alternative causes, coded as the background cause *B,* are equally prevalent regardless of whether *C* is present. The instructions used in our Experiment 1 were intended to encourage participants to make the assumption of random assignment (see also Buehner et al., 2003, Experiments 2–3; Liljeholm & Cheng, 2007). If experimental instructions are unclear on this point, some participants may instead assume that in the make-believe experimental context, when *C* is present, *B* no longer occurs. This alternative assumption would be unwarranted in any realistic situation because it is impossible to know what all the background causes of an effect are, let alone eliminate them. This assumption would imply that the base rate of *E* is irrelevant to assessing the strength of *C* (rendering the very concept of contingency also irrelevant). It would also make it more difficult to detect any effect of SS priors (as generative causes are assumed to compete only when they co-occur) and hence may have contributed to the lack of any advantage for Model I relative to Model II in fitting the meta-analysis data. It is likely that the residual error in the predictions of the Bayesian models (approximately 8% of the variance unaccounted for) in part reflects the vagaries of the instructions and queries employed across the 10 different studies.

Quantitative Comparisons With Nonnormative Models

As argued earlier, the causal query used in the present Experiment 1 is less ambiguous than the rating scales used in most of the previous experiments included in the meta-analysis. Moreover, by mentioning random assignment to conditions, the instructions encouraged the assumption that alternative causes occur independently of the candidate cause. It is therefore instructive to compare the fits of Bayesian and nonnormative models to the data for strength judgments obtained in our Experiment 1. Armed with several free parameters, rules such as EI have an obvious advantage over more parsimonious normative models in fitting any individual data set. The critical test, however, is to what extent nonnormative models robustly generalize to new data sets. The fact that the two recent meta-analyses yielded two different winning nonnormative models is grounds for caution.

As an initial test of the generality of each, we fitted both the EI rule and the H rule to the data from Experiment 1. Using the values of four free parameters as estimated from the meta-analysis data (Perales & Shanks, 2007), the EI rule produced an overall correlation of $r = .95$, slightly lower than that obtained using Bayesian Models I and II based on the power generating function ($r = .98$ for each). In absolute terms, the fit of the EI rule was seriously off (RMS = 22.11, compared with < 8 for either Bayesian model), indicating that the estimated values of its free parameters do not generalize to the present Experiment 1. The rule runs into even worse trouble in explaining the asymmetry between causal directions, predicting an effect of causal direction opposite to that observed in the human data. Because of its free parameters, the EI rule does allow for the possibility that generative and preventive conditions matched on causal power and optimality of the base rate could yield asymmetrical causal strengths. However, for the conditions tested in Experiment 1, the EI rule predicted an overall generative advantage of 10.75 on the 100-point scale, whereas the human data actually showed an overall preventive advantage of 5.98.

Adequacy of fit for the H rule was similar to that for the EI rule. The overall correlation was high, with $r = .96$, but more detailed measures of fit proved problematic. Hattori and Oaksford (2007) did not interpret values of H as a ratio scale of strength but rather

used linear regression to estimate two free parameters (slope and intercept) that optimize its fit to any given data set. After applying linear regression to predict the mean strength judgments obtained in Experiment 1, the H rule gave an RMS value of 18.27, more than double that of the Bayesian models. Moreover, the H rule yields estimates of causal strength that are strictly symmetrical with causal direction; hence, the rule is unable to account for the asymmetries observed in the data from Experiment 1.

The Bayesian and nonnormative models thus proved to be asymmetric in their generalizability: The Bayesian models based on the power generating function can account for a meta-analysis data set that has been used to support nonnormative models (Perales & Shanks, 2007), whereas the leading nonnormative models proved less successful in modeling the present data that support the Bayesian models. Notably, the Bayesian models most clearly show their superiority for a more interpretable data set (namely, Experiment 1), in which some of the methodological problems that affected the meta-analysis data set have been removed. In the General Discussion, we further consider the relative promise of Bayesian versus nonnormative models.

Judgments of Causal Structure

We now assess the SS power model as an account of judgments of causal structure. On the basis of the comparison of Bayesian models for strength judgments, we can confidently reject models based on the linear generating function. Accordingly, we extended Model I (SS power) to account for judgments of causal structure (see Appendix B for mathematical formulation). We compare its performance with the analogous extension of Model II, which is simply the causal support model of Griffiths and Tenenbaum (2005).

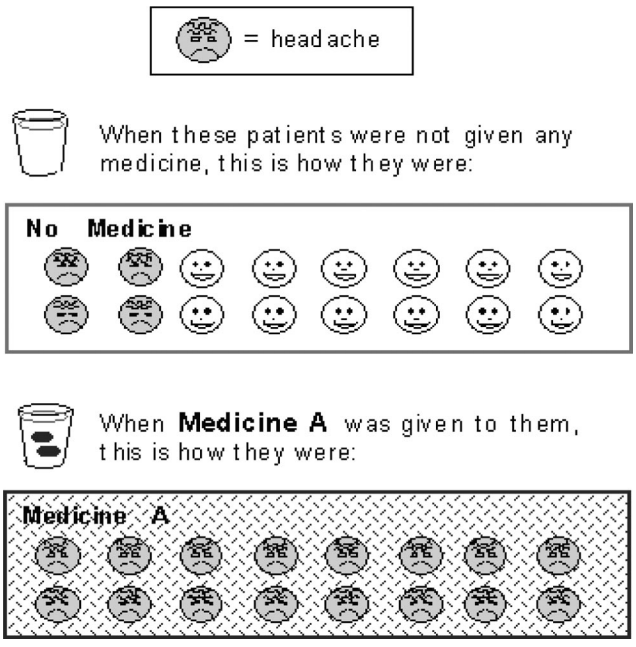

Figure 8. Example of an experimental display used in experiments with the headache cover story, showing patients who had not (top) or had (bottom) received a mineral in an allergy medicine and who either had or had not developed headaches.

Experiment 3: Varying Power, Base Rate, Sample Size, and Direction of Causation

The goal of Experiment 3 was to obtain human judgments about causal structure across a range of matched generative and preventive conditions, enabling tests of alternative Bayesian models. Unlike any of the experiments reported by Griffiths and Tenenbaum (2005), we elicited structure judgments for preventive as well as generative causes while also varying sample size.

Method

Participants. Fifty-three UCLA undergraduates served in the study to obtain partial credit in an Introductory Psychology course.

Materials and procedure. As in Experiment 1, a simultaneous presentation format (see Figure 8) was used to minimize memory demands and other processing issues extraneous to causal inference. Participants first read a cover story about a pharmaceutical company investigating whether various minerals in an allergy medicine might produce headache (generative condition) as a side effect. The preventive cover story was identical except that the word *prevent* was substituted for *produce*. Participants were further informed that each mineral was to be tested in a different lab and that the number of patients who had a headache before receiving any mineral, as well as the total number of patients, would vary across patient groups from different labs. Participants were then presented with data from the tests of the allergy medicine. Each trial was depicted as the face of an allergy patient. As illustrated in Figure 8, each patient was represented by a cartoon face that was either frowning (headache) or smiling (no headache). The data were divided into two subsets, each an array of faces.

The stimuli were presented on a computer and booklets were provided for making causal ratings. Participants were presented with 16 consecutive screens (randomly ordered for each participant). Each screen showed a particular group of patients both before and after they received a mineral. The measure of causal structure was derived from a query asking whether, as a side effect, a mineral in the allergy medicine caused headache (generative conditions) or relieved headache (preventive conditions). Specifically, the query (generative conditions) was "How likely is it that this mineral produces headache?" with the response being a numerical rating on a line marked in units of 10 from 0 (*extremely unlikely*) to 100 (*extremely likely*).[8] For preventive conditions, *produces* was replaced by *relieves*. The dependent measure was the mean rating in each condition.

Design. The design encompassed 32 conditions[9] defined by the factorial combination of eight different contingencies, two sample sizes (16 and 64), and two causal directions (generative and preventive causes). The specific contingency conditions are shown in Figure 9. The contingency conditions and sample size were varied within subjects, while direction of causal influence

[8] T. Griffiths and J. Tenenbaum assisted M. Liljeholm in developing this wording of the structure query (T. Griffiths & J. Tenenbaum, personal communication, January 25, 2005).

[9] Because of a computer error, data were not collected for the preventive condition with a base rate of 0, power of 1, and sample size of 64. This condition and its matched generative condition were therefore dropped from the design. Accordingly, Figure 9 depicts the 30 conditions that were actually used in our analyses.

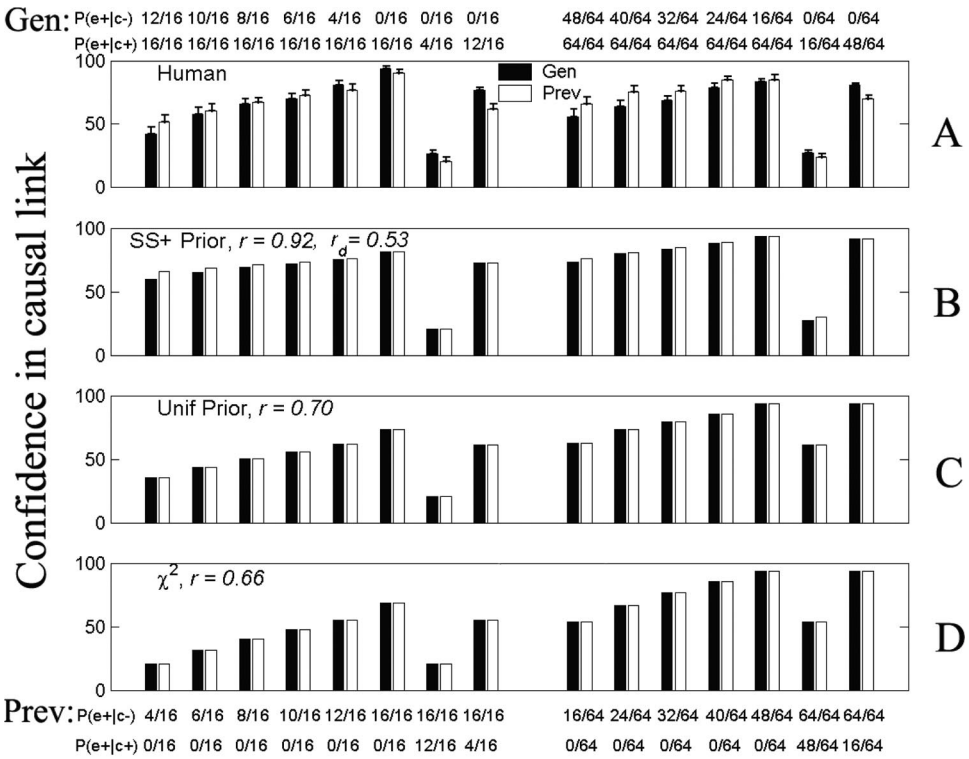

Figure 9. Confidence in a causal link (Experiment 3). Numbers along the top show stimulus contingencies for generative cases; those along the bottom show contingencies for matched negative cases. A: Mean human confidence judgments (error bars indicate one standard error). B: Predictions of the SS power model. C: Predictions of the causal support model. D: Predictions of the chi-square statistic. Gen = generative; Prev = preventive; SS = sparse and strong; Unif = uniform.

was varied between subjects. Twenty-six participants were tested with the generative conditions and 27 with the preventive conditions.

Results

The mean ratings of causal structure for all conditions are shown in Figure 9A. An ANOVA was performed to examine the overall influence of contingency, sample size, and causal direction. Because one contingency condition was missing from the design for sample size 64 (see footnote 9), the matched condition for sample size 16 was omitted from the analysis to allow a factorial ANOVA design. This design included seven contingencies (within-subjects), two samples sizes (within-subjects), and two causal directions (between-subjects). The main effect of contingency was highly significant, $F(6, 306) = 114.6$, $p < .001$. As is apparent from the pattern shown in Figure 9A, the main differences among contingency conditions reflect increased ratings as the causal power increases and decreased ratings as the base rate moves away from the optimal value of 0 (generative conditions) or 1 (preventive conditions). Causal ratings were modestly but reliably higher for the larger than for the smaller sample size (means of 66.9 and 59.1, respectively), $F(1, 51) = 32.1$, $p < .001$. The overall effect of generative versus preventive causes was not reliable ($F < 1$), but causal direction interacted with contingency conditions, $F(6, 306) = 4.76$, $p < .001$. No other interactions approached significance.

To examine the interrelationships among causal direction, base rate, and sample size more carefully, a second ANOVA was performed using just those matched contingencies for which $P(e^+|c^+) = 1$ (corresponding to power = 1) and the base rate $P(e^+|c^-)$ varied systematically from .75 to 0 (generative) or from .25 to .75 (preventive; i.e., the left five contingencies in Figure 9A for each sample size). The effect of base rate was highly reliable, $F(4, 204) = 44.3$, $p < .001$, confirming the clear pattern of lower causal ratings as the base rate departed from the optimal value (0 for generative, 1 for preventive). Overall, ratings were higher for the larger sample size, $F(1, 51) = 30.4$, $p < .001$, with the impact of base rate being reduced for the larger sample size, $F(1, 51) = 5.27$, $p < .05$, for the linear component of the interaction. In addition, as is apparent in Figure 9A, preventive ratings tended to be higher than generative ratings when the base rate was nonoptimal, with the difference diminishing as the base rate approached optimal, $F(1, 51) = 4.37$, $p < .05$, for a test of the monotonic component of the interaction. A separate comparison of the effect of causal direction for the conditions in which the generative base rate was .75 (.25 preventive) versus .25 (.75 preventive) yielded a significant interaction, $F(1, 51) = 4.71$, $p = .035$. As discussed

earlier, these are cases in which the SS power model predicts a preventive advantage.

Comparison of Computational Models

Figure 9 shows the data for human causal judgments (see Figure 9A) along with predictions based on SS+ priors (see Figure 9B), uniform priors (see Figure 9C), and the chi-square statistic (see Figure 9D). For the SS+ priors, the value of α was set to 5 (as in the simulations of strength judgments reported earlier), and the value of β was set to 20 on the basis of a grid search using the human data. As was the case for strength judgments, the SS component of SS+ priors predicts subtle asymmetries between causal judgments across the two causal directions. In contrast, the support model (uniform priors) and chi-square statistic predict that matched generative and preventive contingencies will yield identical mean structure ratings. Accordingly, for the SS+ model only, we computed not only the overall correlation of model predictions with human data but also r_d, the correlation between the observed and predicted difference between mean structure judgments for matched generative and preventive contingencies. (For the support model and chi-square, r_d is not computable because the predicted difference score is always 0.)

As is indicated in Figure 9, the overall correlation was substantially higher using the model based on SS+ priors ($r = .92$, $r_d = .53$) than with uniform priors ($r = .70$) or the chi-square statistic ($r = .66$). Two qualitative aspects of the data favor the model with SS+ priors. First, SS+ priors capture the fact that human judgments of confidence in a causal link are more sensitive to causal power and $P(e^+|c^-)$ (base rate of the effect; e.g., increasingly optimal across left six contingencies in Figure 9) than to sample size. Uniform priors place relatively greater weight on sample size. Indeed, the component of the prior favoring strong causes (i.e., the β parameter in the SS+ priors) by itself yields a correlation of .90 with the observed data. However, in addition, the SS component of SS+ priors captures the apparent asymmetry between generative and preventive judgments for cases matched on causal power (fixed at 1) and optimality of the base rate. For the human data, for 9 of the 10 matched conditions in which the base rate is nonoptimal, the mean preventive rating exceeds the generative case. The asymmetric SS component of SS+ priors captures this subtle difference between preventive and generative judgments, yielding a positive value of r_d (.53, $p < .05$). In contrast, the model with uniform priors and the chi-square statistic predict strict equality of matched generative and preventive conditions.

Are Structure and Strength Judgments Empirically Distinct?

Griffiths and Tenenbaum (2005) applied their causal support model to data from experiments designed to elicit judgments of causal strength, suggesting that people often assess structure when asked to evaluate strength. However, Perales and Shanks (2007) reported that causal support provided a poor overall fit to data from their meta-analysis of causal strength judgments. By comparing performance of Bayesian models for structure versus strength judgments when each is fitted to data based on the alternative type of query, we can assess whether or not these two types of causal queries elicit distinct patterns of data. When fitted to the strength data of Experiment 1, the structure models yielded correlations of $r = .82$ and .80 for SS+ priors and the causal support model, respectively. These fits are much poorer than the correlation of .98 achieved by Models I and II, the most successful strength models. The strength models fared somewhat better when applied to the structure data of Experiment 3, reflecting the fact that SS+ priors are strongly influenced by causal strength (through the β parameter): Model I (SS power) and Model II (uniform) both yielded $r = .86$. Although these fits of strength models to the structure data of Experiment 3 actually surpass the performance of the causal support model, they are notably less adequate than is the fit of the structure model with SS+ priors. These analyses confirm that when the questions are clearly worded, strength and structure queries elicit judgments that are empirically as well as theoretically distinct. Contrary to the tack taken by Griffiths and Tenenbaum, causal support does not provide an adequate model of strength judgments.

Experiment 4: Test of the Influence of Power Versus Sample Size

Experiment 4 was designed to further contrast predictions of the SS power model and the support model assuming uniform priors. The predictions of the two models differ in their sensitivity to sample size versus causal power. The results of Experiment 3 indicated that human support judgments are less influenced by sample size than is predicted by the support model. The present findings are consistent with previous results indicting that human causal judgments are fairly insensitive to sample size when the total number of cases lies in a range similar to that used in Experiment 3 (e.g., Baker et al., 1993; Shanks, 1985; Shanks & Dickinson, 1987). Because it includes substantive priors, the SS power model implies that structure judgments will be less dependent on sample size than is predicted by the support model. In cases when the presented contingencies closely match the SS+ priors, the SS power model predicts that people will be highly confident in the presence of a causal link after only a few observations.

Method

Participants. A total of 107 UCLA undergraduates served in the study to obtain partial credit in an introductory psychology course, with from 24 to 31 participants in each of four conditions. All were tested in a group setting, using booklets that included other experiments.

Materials, design, and procedure. The same basic headache cover story and presentation format were used as in Experiment 3. Four contingency conditions were tested, based on generative causes only. A between-subjects design was used, with each participant evaluating a single problem. Accordingly, rather than referring to multiple minerals within a medicine as in Experiment 3, the cover story simply referred to an allergy medicine that might cause headache as a side effect.

The design compared judgments for two contingencies close to the generative peak for SS+ priors (0/8, 8/8, and 2/8, 8/8) with a small sample size of 8 to two contingencies far from the peak of SS+ priors (0/64, 16/64, and 16/64, 48/64) with a substantially larger sample size of 64 (generative conditions only). The query

was to select one of two alternatives—"This medicine has absolutely no influence on headache"(no link) or "This medicine produces headache"(link exists)—rating confidence in the answer on a 100-point scale. The dependent measure was mean confidence that a link exists (treating the rating as negative when the answer was that no link exists).

Results

Mean confidence in a causal link for each condition is shown in Figure 10A. To test the relative impact of power and sample size on human judgments of confidence in a causal link, a *t* test was performed to compare the average of the two conditions with high power (1.0) but low sample size (left two bars in Figure 10A) with the average of the two conditions with lower power (.25 and .67, respectively) but high sample size (right two bars in Figure 10A). The mean confidence ratings proved to be reliably higher for the former conditions (76.5) than for the latter (51.3), $t(105) = 2.89$, $p < .005$, demonstrating that high power is able to offset low sample size.

Comparison of Computational Models

Model fits provided quantitative confirmation of the ordinal pattern described above. The SS power model (see Figure 10B) yielded a high positive correlation across the four conditions ($r = .82$), whereas the correlations were actually in the wrong direction for both the model with uniform priors ($r = -.16$; see Figure 10C) and chi-square ($r = -.15$; see Figure 10D). People placed much greater weight on match to SS+ priors than on sample size. In the most dramatic case, where the data matched the generative peak at $w_0 = 0$, $w_1 = 1$, human mean confidence was 85 on the 100-point scale after just 16 observations. SS+ priors closely matched the human level of high confidence, whereas uniform priors erroneously predicted a confidence level below 50. Strikingly, uniform priors and chi-square generated the wrong ordinal ranking of this favorable contingency relative to the rightmost condition in Figure 10 (a case of lower power with a much higher sample size).

The Rapidity of Children's Causal Inferences

A basic consequence of generic priors for SS causes is that people are willing to infer a new cause–effect relation on the basis of a limited sample of data when the causal structure matches their priors. The SS power model therefore offers an explanation of how children are able to draw strong causal inferences based on a small number of observed events. For example, Gopnik et al. (2001, Experiment 1) showed 4-year-old children a series of novel toys, some of which were said to be "blickets," which would activate a machine called a "blicket detector" when placed upon it. The instructions to the children strongly implied that the blicket detector was only activated by toys that were blickets ("blickets make the machine go"), not by other background causes. The test question "Is this one a blicket?" probed the existence of a causal link (i.e., structure) rather than causal strength. Gopnik et al. found that 97% of 4-year-old children agreed that a toy block was a blicket if it activated the blicket detector three times in a row and that 78%

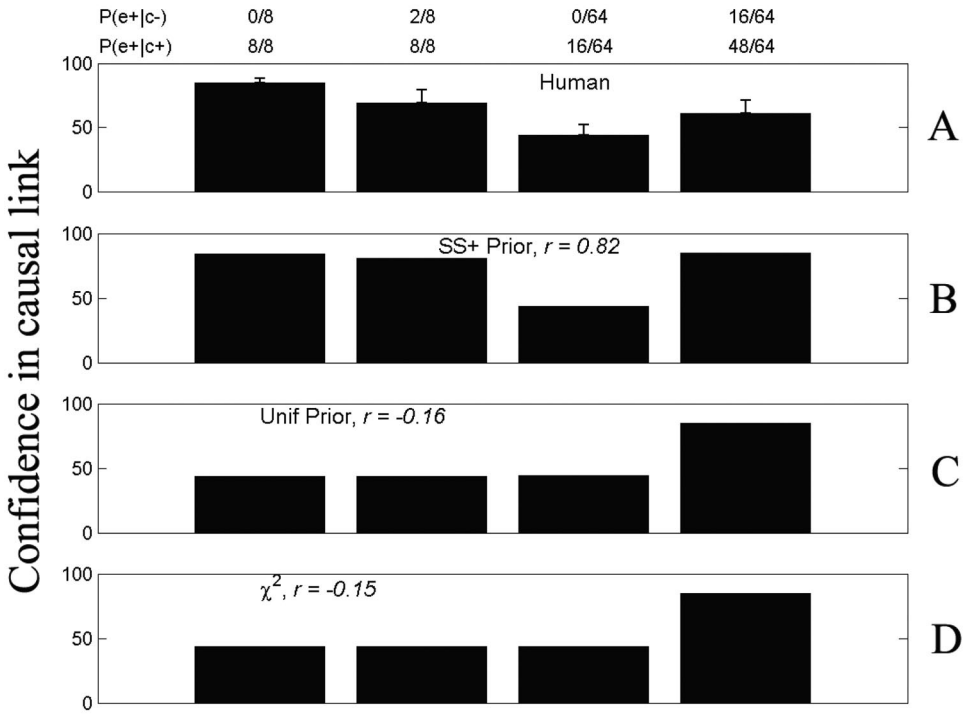

Figure 10. Confidence in a causal link (Experiment 4). A: Mean human confidence judgments (error bars indicate one standard error). B: Predictions of the SS power model. C: Predictions of the causal support model. D: Predictions of the chi-square statistic. SS = sparse and strong; Unif = uniform.

agreed it was a blicket if it activated the detector two out of three times.

In Appendix C, we show how the SS power model of structure judgments captures this rapid causal learning. Our simulation uses generic priors identical to those used in the simulations of experiments with adults, reported above, coupled with a specific prior that the background does not activate the blicket detector (consistent with the instructions that Gopnik et al., 2001, gave to children in their study). Table C1 in Appendix C summarizes predictions derived from the SS power model and other Bayesian models for four experimental conditions. (Additional Bayesian simulations of this data set were reported by Griffiths & Tenenbaum, 2007; see their Table 20-1, p. 335.)

This simulation illustrates how the SS power model generalizes to situations involving multiple potential causes in addition to the constant background, as well as how generic priors can be integrated with a specific prior (in this example, that the strength of the background cause will be 0). Given three trials in which a toy activates the blicket detector, the SS power model yields a high positive support ratio (expressed on a log scale) of 7.48; when just two of three presentations of the potential cause yield the effect, the support ratio is lower but still clearly positive (2.64).

Gopnik et al. (2001) also found that very few children (16%) agreed a block was a blicket when it was paired on two positive trials with another block that had previously activated the detector when presented by itself on a single trial. That is, the presence of a known cause reduced causal learning to a subsequent second cause when the evidence was ambiguous (a version of the well-known phenomenon of blocking, first identified in studies of animal conditioning; Kamin, 1968). In agreement with this finding, the SS power model (which favors a single strong cause) yields a negative support ratio (−2.85) for the blocked cue in Gopnik et al.'s experiment (i.e., the model strongly favors a causal structure in which the blocked cue has no link to the effect). The derivation provided in Appendix C illustrates how the SS power model can be generalized to account for phenomena such as blocking that involve multiple potential causes.

The rapid learning observed in developmental studies such as that of Gopnik et al. (2001; see also Sobel & Kirkham, 2007) disconfirms a variety of alternative models of human causal inference. In particular, it has been proposed that people apply constraint-based algorithms to extract causal structures formalized as Bayes nets. These non-Bayesian Bayes-net models (suggesting a possible excess of terms honoring the Reverend Bayes!) employ the formalism of causal Bayes nets but not Bayesian inference; rather, they rely on data-driven hypothesis testing (Pearl, 1988; Spirtes et al., 2000). Constraint-based models have important practical applications in artificial intelligence, as these algorithms can extract causal networks from masses of contingency data even when cause–effect direction is not established by temporal order or prior knowledge. In knowledge-engineering applications, constraint-based Bayes nets offer valuable supplements to human observers.

At present, however, there is no evidence that these data-intensive, bottom-up algorithms are relevant to psychology. Indeed, constraint-based models constitute valuable tools for knowledge engineering precisely because they do not learn like humans. Human learners have great difficulty extracting cause–effect relations in the absence of critical cues provided by perceived temporal order and their own interventions (Lagnado & Sloman, 2004; Steyvers et al., 2003; see Lagnado et al., 2007). Gopnik et al. (2004) argued that constraint-based learning might somehow account for developmental findings from their blicket paradigm but provided no fits to any experimental data. As reviewed above, these data in fact show that young children draw strong causal conclusions from a handful of observations. To derive these simple causal inferences, the number of observations needed by constraint-based algorithms would exceed that required by human children by two orders of magnitude (as acknowledged by Gopnik et al., 2004, p. 17). Danks (2004) defended the approach but (like Gopnik et al., 2004) provided no fits to any human data. Lacking any theory of priors, constraint-based algorithms are also unable to account for the range of phenomena observed with adult human learners that are the focus of the present article. In contrast, a Bayesian model incorporating generic priors for SS causes is able to explain rapid causal inference on the human scale.

General Discussion

Summary and Implications

We have compared alternative Bayesian models of causal learning as predictors of human judgments of both causal strength and causal structure (existence of a causal link). The central theoretical issues addressed are the form of human priors about causal links and the form of the generating function used by humans to make causal inferences from contingency data. We began with a systematic comparison of predictions derived from Bayesian models that incorporate either the power generating function (Cheng, 1997) or a linear generating function based on ΔP (Jenkins & Ward, 1965; Shanks & Dickinson, 1987). These alternative generating functions were factorially combined with either uniform priors or generic priors for SS causes. Model fits for data from Experiment 1 revealed that models based on the power generating function were considerably more successful overall than those based on the linear generating function. Without any further parameter fitting, the former models also proved quite successful in fitting data sets from a meta-analysis based on 17 experiments selected from 10 studies in the literature (Perales & Shanks, 2007), performing at least as well as the leading nonnormative model (which has four free parameters). By providing a treatment of uncertainty, these Bayesian models can explain phenomena previously viewed as inconsistent with normative models, such as variations in strength estimates with base rate of the effect when the actual contingency is zero.

Models incorporating SS priors were able to account for subtle asymmetries in causal judgments across generative versus preventive causes. We confirmed a novel prediction that people will expect the base rate of the effect to be higher if the candidate cause is described as preventive rather than generative (Experiments 2A and 2B). The SS power model—a Bayesian formulation of causal inference that combines the power generating function with SS priors—provided the best overall account of human strength judgments. The SS power model was extended to create a Bayesian ideal observer model for sequential learning, which was compared with data reported by Shanks (1995). This Bayesian model, which assumes perfect memory for observations, learned more quickly than humans. At a qualitative level, the model accounted for the

standard negatively accelerating learning curve, more positive strength ratings early in learning for a zero-contingency condition when the probability of the effect is relatively high, and also an apparent asymmetry between strength ratings for generative versus preventive causes.

We then extended the SS power model to provide an account of human judgments of causal structure and compared its predictions with those of Griffiths and Tenenbaum's (2005) causal support model. The SS power and support models both incorporate the power generating function, but whereas the support model assumes uniform priors, the SS power model assumes that people prefer causes to be SS. In particular, when making structure judgments, we assume that people are much more willing to infer that a new candidate cause is viable if it has high strength. These generic priors predict a range of phenomena concerning human judgments of causal structure. In Experiments 3–4, the SS power model provided a better quantitative account of human structure judgments than did the support model. The SS power model was able to explain the dominance of power and base rate of the effect over sample size, as well as subtle asymmetries between structure judgments for generative versus preventive causes. In addition, the SS power model provides a qualitative account of rapid causal learning by children (Gopnik et al., 2001).

The failure of alternative Bayesian models based on the linear generating function is especially instructive. Griffiths and Tenenbaum (2005) pointed out that the linear and power generating functions can both be given a Bayesian formulation. A possible misconstrual would be that Bayesian modeling somehow makes the choice of a generating function irrelevant. On the contrary, the Bayesian framework simply derives rational predictions from stated theoretical premises: If a reasoner has certain entering causal beliefs (e.g., that causes independently generate their effects), then some pattern of rational causal judgments follows. The indispensability of theory-based premises in formulating rational models was emphasized by Jaynes (2003), who began his treatise on Bayesian inference with a warning:

> Firstly, no argument is stronger than the premises that go into it, and as Harold Jeffreys noted, those who lay the greatest stress on mathematical rigor are just the ones who, lacking a sure sense of the real world, tie their arguments to unrealistic premises and thus destroy their relevance. (Jaynes, 2003, p. xxvii)

The simulations presented in the present article show that the linear generating function embodied in models based on ΔP, including the asymptotic Rescorla-Wagner model, fails as the basis for a psychological theory of human causal judgments for binary variables—the underlying premise that causes combine in a linear fashion (implying that, by default, causes influence their effect in a mutually exclusive manner rather than independently) turns out to be false. Our analyses confirm the similar conclusions from many earlier studies that did not make use of quantitative Bayesian modeling (e.g., Buehner et al., 2003; Liljeholm & Cheng, 2007; Novick & Cheng, 2004; Wu & Cheng, 1999; for related evidence, see Waldmann & Hagmayer, 2005; Waldmann & Holyoak, 1992; Waldmann, Holyoak & Fratianne, 1995). By providing an explicit model of uncertainty, the Bayesian framework reveals the form of the generating function that guides human causal learning with binary variables.

Comparison of Rational Versus Nonnormative Models

In the present article, we have shown that Bayesian models (specifically, Models I and II based on the power generating function) account for a meta-analysis of causal strength judgments at least as well as do leading nonnormative models in the literature (Perales & Shanks, 2007). In contrast, two leading nonnormative models, the EI rule (Perales & Shanks, 2007) and the H rule (Hattori & Oaksford, 2007), proved less successful in fitting the data from our Experiment 1. Nonetheless, given the longstanding (Schustack & Sternberg, 1981; Ward & Jenkins, 1965) and continuing claims that some nonnormative model can account for human causal judgments, it seems useful to critically examine the plausibility of nonnormative approaches.

We use the term *nonnormative* (following Perales & Shanks, 2007) to refer to models not derived from a well-specified computational analysis of the goals of causal learning. Almost always, the proponents of such models offer some rationale to support claims that the proposed rule is adaptive, efficient, simple, or otherwise plausible as a psychological algorithm. Occasionally the proponents also claim their favored rule is in fact normative (despite the absence of a computational analysis of causal goals), typically on the grounds the rule corresponds to some measure that statisticians have offered to quantify contingency relationships. Given the plethora of nonnormative proposals for assessing causal strength (roughly 40 variants have been proposed; Hattori & Oaksford, 2007), we focus here on the EI and H rules, the apparent winners in the meta-analyses of Perales and Shanks (2007) and Hattori and Oaksford (2007), respectively. (For additional critiques of linear combination rules and various other nonnormative models, see Buehner et al., 2003; Cheng, 1997; Cheng & Novick, 2005; Cheng et al., 2007.)

Two Nonnormative Models

The EI rule, a slight modification of a proposal by Busemeyer (1991), is a somewhat complex variant of the class of linear combination rules, which assign explicit weights to the four cells of the standard 2×2 contingency table. Translating from Bayesian notation to the traditional cell labels, $N(c^+, e^+)$ is the frequency of Cell A, $N(c^+, e^-)$ is the frequency of Cell B, $N(c^-, e^+)$ is the frequency of Cell C, and $N(c^-, e^-)$ is the frequency of Cell D. Qualitatively, high frequencies in Cells A and D tend to confirm a high (generative) strength estimate for the candidate cause *C,* whereas high frequencies in Cells B and C tend to disconfirm a high strength estimate. The EI rule (Perales & Shanks, 2007, p. 583) computes the difference between (a) confirmatory evidence based on frequencies of Cells A and D divided by the total cell frequency and (b) disconfirmatory evidence based on frequencies of Cells B and C divided by the total cell frequency. Critically, the EI rule includes four free parameters corresponding to weights on the four cells. When the EI rule was fitted to the data from their meta-analysis, Perales and Shanks (2007) obtained weight estimates for each cell ordered A > B > C > D (cell weights of .84, .58, .39, and .33, respectively), an empirical ordering commonly observed in studies of contingency learning (see McKenzie & Mikkelsen, 2007). The above parameter values were used in fitting the EI rule to the data from Experiment 1.

The H rule (Hattori & Oaksford, 2007) is characterized as a heuristic simplification of a normative rule, the phi statistic, for

computing the degree of statistical linkage between two binary variables. For binary variables that form a 2×2 contingency table, phi is equivalent to Pearson's *r,* a measure of linear correlation; it corresponds to the chi-square statistic corrected to remove the influence of sample size. Phi is nonnormative by our criterion, as statistics such as chi-square are simply measures of contingency between observable variables. Far from being based on any computational analysis of causal learning, such measures make no reference at all to underlying causal relationships (Cheng, 1997; Cheng et al., 2007). The phi statistic is no more normative as a basis for assessing causal learning than is chi-square.

The H rule is even less normative, differing from the phi statistic in that it ignores the D cell (i.e., cases in which neither the candidate cause nor the effect are present). Hattori and Oaksford (2007) argued that ignoring the frequency of cases in the D cell reduces the burden on memory and will not seriously bias estimates of causal strength as long as the frequency of the D cell is large. Under the extreme rarity assumption—that the base rates of both the cause and the effect are in fact very small—Hattori and Oaksford argued that the H rule is "adaptively rational" (Hattori & Oaksford, 2007, p. 787). They nonetheless acknowledge that ignoring the D cell (and hence the actual base rate of the effect) "is not a norm or a golden rule for causal induction" (Hattori & Oaksford, 2007, p. 772).

Hattori and Oaksford (2007) justified the H rule by claiming that people will be unable to estimate the frequency of the D cell in real-world situations. For example, in assessing whether a fertilizer has a causal link to plant yield (assuming yield to be binary, high or low), the D cell would consist of unfertilized low-yield plants. However, this justification is undermined by the fact that the H rule (unlike the phi statistic) has to be reformulated to deal with preventive causes. Hattori and Oaksford (2007, p. 773) provided a fix for preventive causes that involves swapping the A cell with the B cell and the C cell with the D cell. That is, although the D cell is ignored when the candidate cause is generative, instead the C cell is ignored when the candidate is preventive. To determine the direction of causality to be assessed, the H rule requires an estimate of overall sample size, and given this estimate, knowledge of any three of the four cell frequencies suffices to estimate the fourth.

More specifically, Hattori and Oaksford (2007) suggested that people determine causal direction by comparing the number of cases in which effect *E* is present (the sum of the frequencies of the A and C cells) versus absent (the sum of the frequencies of the B and D cells), thereby assessing whether the occurrence or the nonoccurrence of *E* is more rare. They failed to note that the ability to make this comparison implies that people have enough knowledge to estimate the frequency of the D cell (either directly or by subtracting the other three cell frequencies from the overall sample size). Thus, the fact that the H rule can only decide which cell is to be ignored after determining the causal direction to be assessed contradicts the claim that the rule serves to reduce memory load. Indeed, if people really failed to attend to the frequency of the D cell in the presented data (the basic rationale for the H rule), they would be unable to estimate preventive causal strength, which they evidently can do. Hattori and Oaksford concluded that "the dual factor heuristic can handle the case of a preventive cause after appropriate recontextualization, rephrasing, and swapping the columns of the contingency table" (Hattori & Oaksford, 2007, p. 774). They went on to observe, "recontextualization and reversing the truth value might be viewed as too complex for a fast heuristic for covariation detection" (Hattori & Oaksford, 2007, p. 774). This is indeed a valid concern.

Anomalous Predictions of EI and H Rules

Earlier, we saw that the EI and H rules have difficulty accounting for the data from our Experiment 1. Another way to assess the plausibility of these nonnormative rules as general accounts of causal inference is to examine cases in which the rules generate anomalous predictions. Consider, for example, an experiment in which the frequencies of Cells A through D are 2, 8, 2, and 8, respectively (i.e., the contingency is 0). If we apply the EI rule with its four weight parameters as estimated by Perales and Shanks (2007) using their meta-analysis, the value obtained is $-.11$, which is indeed reasonably close to 0. However, if we change the cell frequencies by multiplying the instances in which the cause is present, keeping the same contingency, the rule would predict that the cause becomes more strongly preventive; moreover, if we analogously change the cell frequencies by multiplying the instances in which the cause is absent, the rule would predict that the cause now becomes generative. For example, suppose we change the cell frequencies to 20, 80, 2, and 8. The EI rule now gives a value of $-.42$, indicating that people will judge the cause to be substantially preventive. If instead we use cell frequencies 2, 8, 20, and 80, the rule predicts that people will judge the cause to be substantially generative, having a positive strength of .39. These predictions seem anomalous. By contrast, the SS power model predicts that (assuming causal direction is unknown) all three of the above conditions will be perceived as weakly preventive (with predicted mean strength values of .24, .23, and .29, respectively).

To test the H rule, consider a further thought experiment. If we provide A–D cell frequencies of 5, 5, 5, and 50, respectively (causal power of .45), the H rule yields a sensible strength estimate, .5. However, if we then change the frequencies to 5, 5, 5, and 5 (0 contingency), the H rule yields the rather extraordinary prediction that judged causal strength will still be .5 (reflecting the rule's tacit assumption that the frequency of Cell D approximates infinity even if the data show that the frequency is equal to 5). People readily judge cases similar to the second to be noncausal (e.g., Buehner et al., 2003, Experiment 2). This example illustrates a general failure of the H rule, in which the extreme rarity assumption acts not as a prior that gracefully yields to empirical evidence but simply as an incorrigible bias.

Rational Versus Nonnormative Models: Summary

Our analysis suggests that rational models of causal learning have much more promise than do nonnormative approaches. Paradoxically, models touted as providing simple yet adaptive heuristics may require multiple free parameters to fit data that can be accounted for by theory-based rational models with fewer or even no free parameters. Moreover, these nonnormative models make anomalous predictions for a variety of additional contingency conditions. In the case of the H rule, the rationale offered for its adaptive value collapses upon more careful examination.

We would argue that further progress in understanding human causal learning requires elevating development of computational

theory above formulation of plausible algorithms. Lacking an underlying computational theory of causal learning, the nonnormative models are unable to generalize beyond the simple case of strength estimates for contingency data involving a single candidate cause—indeed, they do not reliably generalize even across different experiments of this restricted type. Algorithmic rules such as EI and H offer no insight into how people can learn causal strengths when multiple causes co-occur, when causes interact, or when multiple causes and effects are linked within more complex causal models. Nor do they explain how people make causal judgments other than about strength.

Prospects for Bayesian Models of Sequential Learning

It should be emphasized that there is no intrinsic incompatibility between Bayesian models and algorithmic models of causal learning; rather, the two approaches are complementary, addressing different levels of analysis (Marr, 1982). For example, although the specific linear updating rule used in the Rescorla-Wagner model can be rejected as an account of human causal learning with binary variables, other sequential learning models of the same general type (i.e., models that sequentially update strength parameters without assuming comprehensive memory for prior observations) deserve to be more fully explored. Such models have the potential to address phenomena related to order of data presentation (e.g., the difference in magnitude between forward and backward blocking) that lie outside the scope of models based on comprehensive memory for prior observations (such as the present version of the SS power model as applied to the data of Shanks, 1995). As we discussed earlier, Danks and colleagues (Danks, 2003; Danks et al., 2003) developed a sequential model based on the power generating function (Cheng, 1997). Yuille (2005, 2006) demonstrated mathematically that linear and nonlinear variants of sequential learning models can perform ML estimation for a range of different probability models, and Yuille (2006) showed formally how Bayesian models at the computational level can be related to algorithmic models of sequential causal learning.

Recent work has begun to explore sequential learning models that update probability distributions over strength weights, rather than simply point estimates. For example, Dayan and Kakade (2000) developed a sequential model that updates a posterior probability distribution based on the linear generating function. More recently, Lu, Rojas, Beckers, and Yuille (2008) generalized the class of noisy-logical functions (of which the power generating function, based on noisy-OR and noisy-AND-NOT, is a special case) and showed how such functions (e.g., noisy-MAX, which is appropriate when the effect variable is continuous rather than binary) can be used as the basis for sequential updating of strength distributions. The resulting Bayesian model of sequential learning can explain several phenomena (including forward and backward blocking) that have been observed in studies of both human causal learning and classical conditioning with rats.

Contributions of the Bayesian Framework

Beginning with the seminal work of Anderson (1990), Bayesian models have been applied to a wide range of high-level cognitive tasks, including memory retrieval and categorization. The present analyses of causal learning complement recent work applying the Bayesian approach to related forms of informal reasoning (e.g., Hahn & Oaksford, 2007; Oaksford & Chater, 2007).

Two general contributions of the overarching Bayesian framework to the development of psychological theories of cognition deserve emphasis. First, the Bayesian framework provides a systematic way to represent uncertainty. It has long been recognized that causal judgments by humans (and, most likely, other animals) are influenced by factors such as the base rate of the effect, sample size, and confounding, which influence the degree of resulting uncertainty after analyzing a set of data (Cheng & Holyoak, 1995). Griffiths and Tenenbaum (2005; Tenenbaum & Griffiths, 2001) deserve full credit for showing how the Bayesian framework, by introducing representations of probability distributions, provides a formal basis for modeling the degree of uncertainty about causal links. Indeed, as Knill and Pouget (2004) highlighted, "this is the basic premise on which Bayesian theories of cortical processing will succeed or fail—that the brain represents information probabilistically, by coding and computing with probability density functions or approximations to probability density functions" (p. 713). In the domain of causal learning, the Bayesian framework allows the predictive power of theories to move beyond point estimates of parameters such as causal power to estimates of their probability distributions.

Second, the Bayesian framework provides a natural formalism for integrating prior beliefs with likelihoods derived from data to draw inferences. In the case of the SS power model, we start with assumptions about generic priors (for SS causes) and about the generating function for binary causal variables (power generating function); the Bayesian framework is then able to derive detailed quantitative predictions about human causal judgments. The SS power model and the causal support model are in full agreement with respect to their psychological assumptions about how people believe multiple binary causes work together to generate effects: Both models are extensions of the power PC theory (Cheng, 1997) that incorporate a Bayesian formulation of uncertainty. As Perales and Shanks (2007) observed, "the power PC model and the structure-learning models form an interlinked theoretical set that derives from and conforms to normative principles" (p. 582). The SS power model goes beyond both the power PC theory and the causal support model by incorporating a psychological theory of priors, in addition to a theory of the generating function. Both components are required to formulate a successful Bayesian model of causal inference.

Future work will need to evaluate the theoretical claim that the mechanism by which the brain makes causal inferences achieves an approximation to inferences made by Bayesian models. Although Bayesian models may appear complex, there is reason to hope that their computations can be realized in neural systems. (See Satpute et al., 2005, for a neuroimaging study distinguishing causal from associative judgments.) At the most basic level, a continuous probability density function (perhaps represented by a few sample points) can be realized by population coding over a pool of neurons, and computations equivalent to mathematical functions such as integration and convolution can be defined over such neural codes (see Dayan & Hinton, 1996; Knill & Pouget, 2004; Rao, 2007). Indeed, there is reason to conjecture that Bayesian models, which operate on entire probability distributions, may be easier to translate into neural representations than are apparently simpler models that operate on point estimates of probabilities. Moreover, as demonstrated by our quantitative comparison of

Bayesian versus nonnormative models of causal strength judgments, Bayesian models fare well in comparison with nonnormative models that have more free parameters and yet are less robust in generalizing to new data sets. Of course, much work will be required to determine if this optimism (which has guided a great deal of recent work in vision) is borne out in the development of neural models of reasoning.

Generalization to Other Types of Causal Judgments

Role of Simplicity

We have interpreted generic priors favoring SS causes as a special case of a more general preference under conditions of uncertainty for causal models that are simple (Lombrozo, 2007). Indeed, simplicity may be a much broader principle guiding cognitive representations (Chater & Vitányi, 2003). When interpreted as a preference for simpler causal models, SS priors may play an important role not only in causal structure judgments but also in the generation and testing of causal hypotheses. As Mill (1843) noted, in realistic situations, multiple causes often produce a given effect (e.g., cancer may be caused by smoking, inhaling asbestos dust, and exposure to many other substances). In such cases, SS priors may serve to guide a search for some unifying hypothesis, or invariance—a common factor that might be found in all the apparently disparate causal situations (e.g., a type of chemical common to all carcinogens). In an experimental test using novel causal relations, Lien and Cheng (2000) found that people search for a level of causal generalization at which a single cause suffices to explain the presence versus absence of an effect. More generally, it may be possible to relate generic priors applied to specific inference problems to overhypotheses derived from higher level inference problems, integrated within a hierarchical Bayesian model (Kemp, Perfors, & Tenenbaum, 2007).

Multiple Causes and/or Effects

The present article has focused on the simplest possible causal situation, involving only a single candidate cause *C* and a background cause *B* that yield a single effect *E.* However, the SS power model can be generalized to situations involving multiple causes and/or multiple effects. Appendix C illustrates an initial generalization to a multicausal situation in which the phenomenon of blocking arises (see Lu et al., 2008, for a more detailed treatment of blocking within a sequential learning model). Yuille and Lu (2008) showed how Bayesian models can also be generalized to more complex situations in which causal interactions may arise (Liljeholm & Cheng, 2007; Novick & Cheng, 2004). Given that human reasoners clearly operate under capacity limits, we would expect simplicity constraints to play a still greater role in guiding selection of causal models as the number of potential cause–effect relations increases.

Causal Attribution in a Bayesian Model

A major strength of a normative theory of causal learning and inference, such as the power PC theory, is that it generates coherent predictions for a wide range of causal questions (Cheng, 1997; Cheng et al., 2007). Because the SS power model is based on the power PC theory, the former provides a natural Bayesian generalization for all predictions made by the latter. As an illustration of additional Bayesian extensions of the power PC theory, we can consider the important case of judgments of causal attribution (Kelley, 1973). Such judgments have the form: Given that *C* occurs with some probability and that on some particular occasion *E* has occurred, what is the probability that *C* was the cause of *E*'s occurrence? Causal attribution is closely related to diagnostic inference—for example, using data about patterns of effects to infer the states of unobserved potential causes (Pearl, 1988; Waldmann & Holyoak, 1992; Waldmann et al., 1995). Causal models code knowledge in terms of influences directed from causes to their effects; however, Bayes' rule provides the basic inference tool required to make inferences that go against the causal arrow, using knowledge of effects to infer the states of their causes.

Some studies that nominally investigated judgments of causal strength used queries that may have elicited causal attribution judgments, at least from some participants. For example, White (2003; included by Perales and Shanks, 2007, in their meta-analysis of data on strength judgments) asked participants to rate "the extent to which [a substance] causes allergic reactions in the patient" (see White, 2003, p. 714). As noted by Cheng and Novick (2005, p. 700), the extent of a cause is a question about causal attribution: That is, given that *E* has occurred, what is the probability that *C* caused it? Cheng and Novick showed, using the power PC theory, that causal attribution (unlike causal strength) is normatively sensitive to prevalence of the cause (e.g., smoking may cause cancer to a great extent if people in the target population commonly smoke, even if the causal power of smoking with respect to cancer is low).

The causal attribution question requires apportioning the observed probability of an effect, $P(e^+)$, among causes of *E.* On the basis of the same assumptions of the power PC theory that we have used throughout the present article, Cheng and Novick (2005, p. 700, Equation 3) derived the predicted probability that *C* is the cause of *E* when *E* occurs, namely,

$$P(c^+ \rightarrow e^+ | e^+) = P(c^+) q_c / P(e^+), \tag{16}$$

where $c^+ \rightarrow e^+$ denotes that *C* is the cause of *E*'s occurrence (corresponding to an unobservable state in a causal model; Cheng, 1997).

Analogous to the basic equations for causal power (see Equations 4–5), Equation 16 yields a point estimate of causal attribution judgments. Analogous to the derivation of the posterior distribution of causal strength (see Appendix A, Equation A1), a Bayesian model can incorporate a likelihood function and priors on causal strength to derive the probability distribution of causal attribution. Statistical quantities computed from the estimated distribution (e.g., the mean, mode, or median) can then be compared with human performance in judgments of causal attribution. This derivation can be readily calculated using the generalized noisy-logical representation (Yuille & Lu, 2008), which includes hidden nodes representing occasions on which some particular factor is the (unobservable) cause of *E*. More generally, for any point estimate of a theoretical quantity derivable from the power PC theory, a corresponding Bayesian estimate can be derived, incorporating a treatment of uncertainty and allowing for the addition of a theory of priors.

Forming Causal Hypotheses

By itself, Bayesian inference addresses reasoning about causal models, including both model selection and strength estimation,

but does not explain how causal hypotheses are formulated in the first place. Of course, if a hypothesis is not represented, no inferences about it are possible. As Jaynes (2003) put it, "if we hope to detect any phenomenon, we must use a model that at least allows for the *possibility* that it exists" (p. xxvi). There are several approaches to modeling the formation of causal hypotheses. For example, Griffiths and Tenenbaum (2007) suggested that hypothesis generation can be modeled by some kind of causal grammar. A complementary possibility, with a long history in the philosophy of science (Hesse, 1966) and psychology (Gick & Holyoak, 1980; Holyoak & Thagard, 1995; Hummel & Holyoak, 2003), is that causal models for novel situations can be generated by analogy to models of situations that are better understood (see Lee & Holyoak, 2008). The mechanisms by which causal hypotheses are formed clearly require additional investigation.

Other Generic Priors

Finally, the concept of generic priors can potentially be generalized to other types of learning. One obvious candidate is category learning. The first psychological model of categorization based on the Bayesian framework, proposed by Fried and Holyoak (1984), simply assumed that perceptual categories are learned by updating the mean and variance of a multidimensional normal distribution. Flannagan, Fried, and Holyoak (1986) extended this notion by proposing that people have priors for the abstract form of the distributions of quantitative dimensions—priors that favor learning of categories with unimodal and symmetrical distributions. Flanagan et al. demonstrated that learning a category that violated this distributional form was relatively difficult but facilitated subsequent learning of a second category (based on different perceptual dimensions) that also violated the unimodal and symmetrical prior. Such evidence suggests that perceptual category learning may indeed be guided by some type of generic priors and that these priors adapt to learning experiences. Armed with recent advances in computational tools for representing priors over probability distributions, it may be possible to provide deeper insights into the role of generic priors across a variety of different types of learning.

References

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 15,* 147–149.

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgments of influence. *Learning and Motivation, 14,* 381–405.

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

Baker, A. G., Mercier, P., Vallée-Tourangeau, F., Frank, R., & Pan, M. (1993). Selective associations and causality judgments: The presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 414–432.

Barlow, H., & Tripathy, S. P. (1997). Correspondence noise and signal pooling in the detection of coherent visual motion. *Journal of Neuroscience, 17,* 7954–7966.

Buehner, M. J., & Cheng, P. W. (2005). Causal learning. In K. J. Holyoak & R. G. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 143–168). Cambridge, England: Cambridge University Press.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1119–1140.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). San Diego, CA: Academic Press.

Busemeyer, J. R. (1991). Intuitive statistical estimation. In N. H. Anderson (Ed.), *Contributions to information integration theory* (pp. 187–205). Hillsdale, NJ: Erlbaum.

Busemeyer, J. R., Myung, I. J., & McDaniel, M. A. (1993). Cue competition effects: Empirical tests of adaptive network learning models. *Psychological Science, 4,* 190–195.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Science, 7,* 19–22.

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

Cheng, P. W. (2000). Causality in the mind: Estimating contextual and conjunctive causal power. In F. Keil & R. Wilson (Eds.), *Cognition and explanation* (pp. 227–253). Cambridge, MA: MIT Press.

Cheng, P. W., & Holyoak, K. J. (1995). Complex adaptive systems as intuitive statisticians: Causality, contingency, and prediction. In H. L. Roitblat & J. A. Meyer (Eds.), *Comparative approaches to cognitive science* (pp. 271–302). Cambridge, MA: MIT Press.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition, 40,* 83–120.

Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and Luhmann and Ahn (2005). *Psychological Review, 112,* 694–707.

Cheng, P. W., Novick, L. R., Liljeholm, M., & Ford, C. (2007). In M. O'Rourke (Ed.), *Topics in contemporary philosophy: Vol. 4. Explanation and causation* (pp. 1–32). Cambridge, MA: MIT Press.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology, 47,* 109–121.

Danks, D. (2004). Constraint-based human causal learning. In M. Lovett, C. Schunn, C. Lebiere, & P. Munro (Eds.), *Proceedings of the 6th International Conference on Cognitive Modeling* (pp. 342–343). Mahwah, NJ: Erlbaum.

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 67–74). Cambridge, MA: MIT Press.

Dayan, P., & Hinton, G. E. (1996). Varieties of Helmholtz machine. *Neural Networks, 9,* 1385–1403.

Dayan, P., & Kakade, S. (2000). Explaining away in weight space. In T. K. Leen, T. G. Diettrich, & V. Tresp (Eds.), *Advances in neural information processing systems* (Vol. 13, pp. 451–457). Cambridge, MA: MIT Press.

Flannagan, M. J., Fried, L. S., & Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 241–256.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 234–257.

Gelman, S., & Kremer, K. E. (1991). Understanding natural causes: Children's explanations of how objects and their properties originate. *Child Development, 62,* 396–414.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306–355.

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

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 1–30.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37,* 620–629.

Graham, D. J., & Field, D. J. (2007). Sparse coding in the neocortex. In

J. H. Kaas & L. A. Krubitzer (Eds.), *Evolution of the nervous system: Vol. 3. Mammals* (pp. 181–197). San Diego, CA: Academic Press.

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

Griffiths, T. L., & Tenenbaum, J. B. (2007). Two proposals for causal grammars. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 323–345). Oxford, England: Oxford University Press.

Grimes, D. B., & Rao, R. P. N. (2004). Bilinear sparse coding for invariant vision. *Neural Computation, 17,* 47–73.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review, 114,* 704–732.

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science, 31,* 765–814.

Hesse, M. (1966*). Models and analogies in science.* Notre Dame, IN: University of Notre Dame Press.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought.* Cambridge, MA: MIT Press.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220–264.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge, England: Cambridge University Press.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs*: *General and Applied, 79*(1, Whole No. 594).

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami Symposium on the Prediction of Behavior, 1967: Aversive stimulation* (pp. 9–31). Coral Gables, FL: University of Miami Press.

Kao, S.-F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1363–1386.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist, 28,* 107–128.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10,* 307–321.

Kim, J. (1993). Causes and events: Mackie on causation. In E. Sosa & M. Tooley (Eds.), *Causation* (pp. 60–74). Oxford, England: Oxford University Press.

Kittur, A., Holyoak, K. J., & Hummel, J. E. (2006). Using ideal observers in higher-order human category learning. In R. Sun & N. Miyake (Eds.), *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 435–440). Mahwah, NJ: Erlbaum.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27,* 712–719.

Kushnir, T., Gopnik, A., Schulz, L. E., & Danks, D. (2003). Inferring hidden causes. In R. Alterman & David Kirsh (Eds.), *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (pp. 699–703). Mahwah, NJ: Erlbaum.

Lagnado, D. A. (1994). *The psychology of explanation: A Bayesian approach.* Unpublished master's thesis, University of Birmingham, Birmingham, England.

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

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

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

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs, 13,* 455–476.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology, 40,* 87–137.

Liljeholm, M. (2007). Structure learning, parameter estimation and causal assumptions. *Dissertation Abstracts International, 68*(1), 643B.

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

Liljeholm, M., & Cheng, P. W. (in press). The influence of virtual sample size on confidence and causal strength judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Lober, K., & Shanks, D. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review, 107,* 195–212.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55,* 232–257.

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. L. (2008). Sequential causal learning in humans and rats. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 185–188). Austin, TX: Cognitive Science Society.

Lu, H., & Yuille, A. L. (2006). Ideal observers for detecting motion: Correspondence noise. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), *Advances in neural information processing systems* (Vol. 18, pp. 827–834). Cambridge, MA: MIT Press.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2006). Modeling causal learning using Bayesian generic priors on generative and preventive powers. In R. Sun & N. Miyake (Eds.), *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 519–524). Mahwah, NJ: Erlbaum.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2007). Bayesian models of judgments of causal strength: A comparison. In D. S. McNamara & G. Trafton (Eds.), *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society* (pp. 1241–1246). Austin, TX: Cognitive Science Society.

Mackay, D. (2003). *Information theory, inference and learning algorithms.* Cambridge, England: Cambridge University Press.

Mackie, J. L. (1974). *The cement of the universe.* Oxford, England: Oxford University Press.

Mamassian, P., & Landy, M. S. (1998). Observer biases in the 3D interpretation of line drawings. *Vision Research, 38,* 2817–2832.

Mandel, D. R., & Lehman, D. R. (1998). Integration of contingency information in judgments of cause, covariation, and probability. *Journal of Experimental Psychology: General, 127,* 269–285.

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

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology, 54,* 33–61.

Mill, J. S. (1843). *System of logic.* London: John Parker.

Newton, I. (1968). The rules of reasoning in philosophy. In A. Motte (Trans.), *The mathematical principles of natural philosophy* (Vol. 2, pp. 202–205). London: Dawson. (Original work published 1729)

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111,* 455–485.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* Oxford, England: Oxford University Press.

Olshausen, B. A., & Field, D. J. (1996, June 13). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381,* 607–609.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcom-

plete basis set: A strategy employed by V1? *Vision Research, 37,* 3311–3325.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology, 14,* 481–487.

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

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

Peirce, C. S. (1931–1958). *Collected papers.* Cambridge, MA: Harvard University Press.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review, 14,* 577–596.

Rao, R. P. N. (2007). Neural models of Bayesian belief propagation. In K. Doya, S. Ishi, A. Pouget, & R. P. N. Rao (Eds.), *The Bayesian brain: Probabilistic approaches to neural coding* (pp. 239–258). Cambridge, MA: MIT Press.

Reichenbach, H. (1956). *The direction of time.* Berkeley: University of California Press.

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

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton, NJ: Princeton University Press.

Satpute, A. B., Fenker, D. B., Waldmann, M. R., Tabibnia, G., Holyoak, K. J., & Lieberman, M. D. (2005). An fMRI study of causal judgments. *European Journal of Neuroscience, 22,* 1233–1238.

Saxe, R., Tenenbaum, J. B., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10- and 12-month infants. *Psychological Science, 16,* 995–1001.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development, 77,* 427–442.

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General, 110,* 101–120.

Shanks, D. R. (1985). Continuous monitoring of human contingency judgment across trials. *Memory & Cognition, 13,* 158–167.

Shanks, D. R. (1987). Acquisition functions in contingency judgment. *Learning and Motivation, 18,* 147–166.

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 48*(A), 257–279.

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

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 64,* 99–118.

Sobel, D. M., & Kirkham, N. Z. (2007). Bayes nets and babies: Infants' developing statistical reasoning abilities and their representation of causal knowledge. *Developmental Science, 10,* 298–306.

Sober, E. (2002). What is the problem of simplicity? In A. Zellner, H. Keuzenkamp, & M. McAleer (Eds.), *Simplicity, inference, and modeling* (pp. 13–32). Cambridge, England: Cambridge University Press.

Sober, E. (2006). Parsimony. In S. Sarkar & J. Pfeifer (Eds.), *The philosophy of science: An encyclopedia* (pp. 531–538). New York: Routledge.

Sosa, E., & Tooley, M. (1993). Introduction. In E. Sosa & M. Tooley (Eds.), *Causation* (pp. 1–32). Oxford, England: Oxford University Press.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd rev. ed.). Cambridge, MA: MIT Press.

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

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems* (Vol. 13, pp. 59–65). Cambridge, MA: MIT Press.

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

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

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

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1102–1107). Mahwah, NJ: Erlbaum.

Ward, W., & Jenkins, H. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology, 19,* 231–241.

Wasserman, E. A., Elek, S. M., Chatlosh, D. C., & Baker, A. G. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 174–188.

Wasserman, E. A., Kao, S.-F., Van Hamme, L. J., Katagiri, M., & Young, M. E. (1996). Causation and association. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 207–264). San Diego, CA: Academic Press.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience, 5,* 598–604.

White, P. A. (2003). Making causal judgments from the proportion of confirming instances: The *pCI* rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 710–727.

White, P. A. (2004). Causal judgment from contingency information: A systematic test of the *pCI* rule. *Memory & Cognition, 32,* 353–368.

Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science, 10,* 92–97.

Yuille, A. L. (2005). The Rescorla-Wagner algorithm and maximum likelihood estimation. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 17, pp. 1585–1592). Cambridge, MA: MIT Press.

Yuille, A. L. (2006). Augmented Rescorla-Wagner and maximum likelihood estimation. In B. Schölkopf, J. Platt, & Y. Weiss (Eds.), *Advances in neural information processing systems* (Vol. 18, pp. 1561–1568). Cambridge, MA: MIT Press.

Yuille, A. L., & Grzywacz, N. M. (1988, May 5). A computational theory for the perception of coherent visual motion. *Nature, 333,* 71–74.

Yuille, A. L., & Lu, H. (2008). The noisy-logical distribution and its application to causal inference. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20, pp. 1673–1680). Cambridge, MA: MIT Press.

(*Appendixes follow*)

Appendix A

Derivation of Bayesian Models of Strength Judgments

On the basis of observation of contingency data *D,* a Bayesian model is able to assess the probability distribution of causal strength w_1 so as to quantify statistical properties of the causal strength of candidate cause *C* to produce or prevent *E*. The posterior distribution $P(w_1|D)$ is obtained by applying Bayes' rule,

$$P(w_1|D,Graph1)=\int_0^1 P(w_0,w_1|D,Graph1)dw_0$$

$$=\int_0^1 \frac{P(D|w_0,w_1,Graph1)P(w_0,w_1|Graph1)}{P(D)}dw_0, \quad \text{(A1)}$$

where $P(D|w_0,\ w_1,\ Graph1)$ is the likelihood term. $P(w_0,\ w_1|Graph1)$ gives prior probabilities that model the learner's beliefs about the values of causal strengths. $P(D)$ is the normalizing term, denoting the probability of obtaining the observed data. The likelihood term $P(D|w_0,\ w_1,\ Graph1)$ is given by

$$P(D|w_0,w_1,Graph1) = \begin{pmatrix} N(c^-) \\ N(e^+,c^-) \end{pmatrix} \times \begin{pmatrix} N(c^+) \\ N(e^+,c^+) \end{pmatrix} \prod_{e,c} P(e|b,c;w_0,w_1)^{N(e,c)}, \quad \text{(A2)}$$

where b,c,e $\in \{0,1\}$ denotes the absence and the presence of the causes *B, C,* and the effect *E,* and $\binom{n}{k}$ denotes the number of ways of picking *k* unordered outcomes from *n* possibilities. $N(c^+)$ indicates the count of events in which the candidate cause is present, with analogous definitions for the other $N(.)$ terms. Figure 3 in the main text shows an example of the posterior distribution of w_1 given contingency data of $p(e^+|c^-) = 12/16$ and $p(e^+|c^+) = 16/16$. In our simulations, we compare the average human strength rating for a given contingency condition with the mean of w_1 computed using the posterior distribution. The mean of w_1 is determined by

$$\bar{w}_1=\int_0^1 w_1P(w_1|D,Graph1)dw_1. \quad \text{(A3)}$$

We implemented four Bayesian models, defined by the factorial combination of two generating functions (linear and power) and two priors (uniform and sparse and strong [SS]). Griffiths and Tenenbaum (2005) showed that causal power (*q* in Equations 4–5 in the main text) corresponds to the maximum likelihood (ML) estimate for the random variable w_1 on a fixed graph (see Graph 1 in Figure 1 in the main text) under the power generating function (see Equations 2–3 in the main text). Because the power generating function obeys the laws of probability, the weights w_0 and w_1 are inherently constrained to the range (0, 1). An alternative generating function that provides a measure of causal strength can be derived from ΔP (see Equation 6 in the main text), which yields a linear generating function,

$$P(e^+|b,\ c;\ w_0,\ w_1) = w_0b + w_1c, \quad \text{(A4)}$$

where w_0 is within the range (0, 1) and w_1 is within the range (−1, 1), and with an additional constraint that $w_0 + w_1$ must lie in the range (0, 1) so as to result in a legitimate probability distribution. Equation A4 simply states that the candidate cause *C* changes the probability of *E* by a constant amount regardless of the presence or absence of other causes, such as *B*. Griffiths and Tenenbaum (2005) proved that Equation A4 yields ΔP (see Equation 6 in the main text) as the ML estimate of w_1 when substituted for Equations 2–3 in the Bayesian model.

The second conceptual component in Equation A1 is the prior on causal strength, $P(w_0,\ w_1)$, within the causal structure of Graph 1 in Figure 1 in the main text. When *C* is an unfamiliar cause, a natural assumption is that people will have no substantive priors about the values of w_0 and w_1, modeled by priors that are uniform over the range (0, 1) (Griffiths & Tenenbaum, 2005). An alternative possibility is that people apply SS generic priors (see Equations 10–11 in the main text) to make strength judgments. For the models with SS priors, we set $\alpha = 5$ after exploring the parameter space in an initial data set (see Lu et al., 2007).

Appendix B

Formalization of SS Power Model for Structure Judgments

In relation to the graphs in Figure 1, a structure judgment involves a relative judgment between Graph 1 and Graph 0. The sparse and strong (SS) power model, like the causal support model, assumes the power generating function (see Equations 2–3 in the main text) and then derives a measure of confidence in a causal link (analogous to support) using Equations 7–9 (see the main text). In the generative case, both causes, *B* and *C,* produce the effect. If data *D* are summarized by contingencies *N*(*e, c*), the number of cases for each combination of presence versus absence of the effect and cause, then the likelihood given causal strengths (w_0, w_1) within the causal structure of Graph 1 is

$$P(D|w_0,w_1,gen,Graph1) = \begin{pmatrix} N(c^-) \\ N(e^+,c^-) \end{pmatrix} \times \begin{pmatrix} N(c^+) \\ N(e^+,c^+) \end{pmatrix} w_0^{N(e^+,c^-)}(1-w_0)^{N(e^-,c^-)}[w_0 + w_1 - w_0w_1]^{N(e^+,c^+)}[1-w_0-w_1 + w_0w_1]^{N(e^-,c^+)}. \quad \text{(B1)}$$

Similarly, the likelihood within the causal structure of Graph 0 (setting $w_i = 0$) is

$$P(D|w_0,Graph0) = \begin{pmatrix} N(c^-) \\ N(e^+,c^-) \end{pmatrix} \times \begin{pmatrix} N(c^+) \\ N(e^+,c^+) \end{pmatrix} w_0^{N(e^+,c^-)+N(e^+,c^+)}(1-w_0)^{N(e^-,c^-)+N(e^-,c^+)}. \quad \text{(B2)}$$

In the preventive case, background cause *B* is again assumed to be generative (Assumption 2 of the power PC theory), hence only *C* could be a preventer (i.e., *B* and *C* do not compete). The likelihood term for Graph 1 is given by

$$P(D|w_0,w_1,prev,Graph1) = \begin{pmatrix} N(c^-) \\ N(e^+,c^-) \end{pmatrix} \times \begin{pmatrix} N(c^+) \\ N(e^+,c^+) \end{pmatrix} w_0^{N(e^+,c^-)}(1-w_0)^{N(e^-,c^-)}[w_0(1 - w_1)]^{N(e^+,c^+)}[1-w_0(1-w_1)]^{N(e^-,c^+)}. \quad \text{(B3)}$$

The likelihood for Graph 0 is the same as in the generative case (see Equation B2)

The second component in Equation 9 in the main text is the prior on causal strength, $P(w_0, w_1|Graph1)$, within the causal structure of Graph 1. Griffiths and Tenenbaum (2005) assumed that the priors on weights w_0 and w_1 follow a uniform distribution. We assume that for structure judgments, people adopt SS+ priors, as defined in Equation 13 in the main text. The value of β was set to 20 in all reported simulations.

For both the generative and preventive cases, $P(w_0|Graph0)$ is obtained as the marginal distribution of $P(w_0, w_1|Graph1)$ by integrating out w_1. Using the marginal distribution of $P(w_0, w_1|Graph1)$ to assign priors on w_0 in Graph 0 ensures that Graph 1 differs from Graph 0 solely by the addition of w_1, without any confounding by a change in priors on w_0 (Jaynes, 2003, p. 612).

$$P(w_0|gen,Graph0) \propto e^{-\alpha w_0} + e^{-\alpha(1-w_0)}. \quad \text{(B4)}$$

$$P(w_0|prev,Graph0) \propto e^{-\alpha(1-w_0)}. \quad \text{(B5)}$$

Appendix C

Simulation of Rapid Causal Learning by Children

We simulated results reported by Gopnik et al. (2001, Experiment 1). The design, together with results for 4-year-old children and three Bayesian models, is summarized in Table C1. A and B refer to toys that may be "blickets," said to cause a reaction in a "blicket detector."

In addition to the generalization of the sparse and strong (SS) power model, two alternative Bayesian models were considered. For all models, we assumed the decision involves four causal graphs, all including a background cause B_G. Graph 0 includes only B_G; Graph A adds *A* as a cause; Graph B adds *B* as a cause; Graph AB adds both *A* and *B* as causes. The support ratio for any potential cause is defined as the log of the ratio of the summed posterior probabilities of graphs that include the relevant cause to those of the graphs that do not include the cause. Thus, the support ratio for cause *A* is

$$\text{support } A = \log \frac{P(D|Graph\ A) + P(D|Graph\ AB)}{P(D|Graph0) + P(D|Graph\ B)}. \quad \text{(C1)}$$

The support ratio for *B* is defined analogously.

For the SS+ power model, the generic priors for Graph AB are

$$P(w_0,w_A,w_B|gen,GraphAB) \propto e^{-\beta(1-w_A)}e^{-\beta(1-w_B)} \times (e^{-\alpha(1-w_0)-\alpha w_A-\alpha w_B} + e^{-\alpha w_0-\alpha(1-w_A)-\alpha w_B} + e^{-\alpha w_0-\alpha w_A-\alpha(1-w_B)}), \quad \text{(C2)}$$

where w_0, w_A, w_B indicate the causal strength of background, cause *A,* and cause *B,* respectively. The parameters α, β, are set to the same values as used in our previous simulations of structure judgments (5 and 20, respectively). The SS+ prior in the three-cause graph is a natural generalization of the prior (see Equation

(*Appendix continues*)

Table C1
Causal Structure Judgments by 4-Year-Old Children (Data From Gopnik et al., 2001, Table 1) and Predictions of Three Bayesian Models

	Children (% "yes')		SS+ with specific prior		Specific prior only		Uniform prior	
Experimental condition	*A*	*B*	*A*	*B*	*A*	*B*	*A*	*B*
Two-cause condition: *A*+ (3), *B*−; *B*+ (2)	97	78	7.48	2.64	5.48	2.91	0.82	−0.21
One-cause condition: *A*+, *B*−; *AB*+ (2)	91	16	6.33	−2.85	4.36	−0.42	1.26	−0.44

Note. In the two-cause condition, order of *A* and *B* trials was counterbalanced; in the one-cause condition, order of *A*-only and *B*-only trials was counterbalanced. SS = sparse and strong.

13 in the main text) in the two-cause graph (see Figure 1 in the main text). The first two exponential terms in Equation C2 model the question-induced prior favoring strong candidate causes ($w_A = 1$, $w_B = 1$), with the same definition as in Equation 12 in the main text. The last term is the generic prior, favoring SS causes, with three peaks in the prior distribution: $(w_0, w_A, w_B) = (1, 0, 0), (0, 1, 0), (0, 0, 1)$. In addition, because participants in the experiment were explicitly informed that the blicket detector never reacts in the absence of a blicket, a specific prior that the causal strength of w_0 is preferred to be 0 is introduced by

$$P(w_0) \propto e^{-\chi w_0}. \quad \text{(C3)}$$

Equation C3 is analogous to Equation 13, with parameter χ set to the same value as β (i.e., 20) to model the experiment-induced prior knowledge before observing any experimental data. Likelihoods for all graphs are defined using the noisy-OR generating function (see Equation 2 in the main text).

For comparison, we also derived predictions from a model with uniform generic priors but the same specific preference that the strength of the background w_0 is 0 (specific prior only) and a model with entirely uniform priors. As shown in Table C1, both the SS power model and the other model with the specific prior capture children's rapid learning that *A* is a cause in the two-cause condition (high positive support ratio) and rejection of *B* as a cause in the one-cause condition (negative support ratio); however, the SS power model gives a greater differentiation between these two extreme conditions. The model with entirely uniform priors is clearly inadequate as an account of the children's data, as in two experimental conditions, it predicts a trend opposite to that observed for human ratings of *A*.

Received March 13, 2007
Revision received June 11, 2008
Accepted June 11, 2008 ■