Modeling Causal Learning Using Bayesian Generic Priors on Generative and Preventive Powers

Hongjing Lu (hongjing@ucla.edu)
Department of Statistics, UCLA, Box 951563, Los Angeles, CA 90095-1563, USA

Alan Yuille (yuille@stat.ucla.edu)
Department of Statistics, UCLA

Mimi Liljeholm (mlil@ucla.edu)
Department of Psychology, UCLA

Patricia W. Cheng (cheng@lifesci.ucla.edu)
Department of Psychology, UCLA

Keith J. Holyoak (holyoak@lifesci.ucla.edu)
Department of Psychology, UCLA

Abstract

We present a Bayesian model of causal learning that incorporates generic priors on distributions of weights representing potential powers to either produce or prevent an effect. These generic priors favor *necessary and sufficient* causes. Across three experiments, the model explains the systematic pattern of human judgments observed for questions regarding support for a causal link, for both generative and preventive causes.

Keywords: causal learning; Bayesian inference

Causal Inference in a Bayesian Framework

Intelligent behavior in a complex and potentially hostile environment depends on acquiring and exploiting knowledge of "what causes what." It is likely that the cognitive mechanisms for causal learning have deep evolutionary roots, a conjecture supported by many parallels between phenomena in animal conditioning and human causal learning (see Shanks, 2004). Ever since the philosopher David Hume, the fundamental question about causal knowledge has been how a learner can take non-causal inputs (notably, observations regarding temporal order and covariation) and induce cause-effect relations as outputs. Cheng (1997) developed a theory that integrates the Humean covariational view of causality with Kant's conception of causal "powers". Her power PC theory assumes that learners have a tacit understanding that causes in the world have powers (i.e., strengths) to produce or prevent effects, and use observations to infer unobservable causal powers (for a review see Cheng et al., in press).

The view that learners have a tacit theory of causal powers can be incorporated into a Bayesian framework for inference. Griffiths and Tenenbaum (2005) developed a Bayesian model, closely related to the power PC theory, for inferring whether a causal link exists between cause *C* and effect *E* (i.e., model selection for the structure of the causal graph; Mackay, 2003). Their model addressed the simplest variant of elemental causal induction, in which the learner is using observations to decide between Graph 0 versus Graph 1 (Fig. 1), where *B* is a constantly-present background cause that may generate *E*, and *C* is a candidate cause that may be either present or absent (varying from trial to trial).

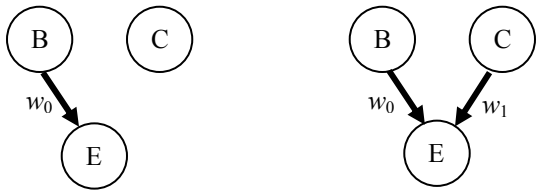

Figure 1. Graphs contrasting hypotheses that *C* causes *E* (Graph 1) or does not (Graph 0). *B*, *C*, and *E* are binary variables. Weights w_0 and w_1 indicate causal strength of the background cause (*B*) and the candidate cause (*C*), respectively.

A major strength of Bayesian inference is that it enables beliefs to be updated by integrating prior beliefs with new observations. Bayesian inference involves two basic components, *likelihood* probabilities and *prior* probabilities. Likelihoods assess the probability that particular observed data would be expected under some hypothesis, and are determined by the generating model for the data (e.g., how multiple independently-operating causes produce an effect). Priors assess beliefs about the world held before observing any particular data (e.g., beliefs about causal powers).

One variant of the "causal support" model developed by Griffiths and Tenenbaum (2005) used a generating model proposed by Cheng (1997), based on a logical "noisy-OR" function (Eq. 4) for generative causes and "noisy-AND-NOT" (Eq. 5) for preventive causes. (See Glymour, 2001, for a more general definition of what he termed "Cheng

models".) This causal-support variant yields causal power (Cheng, 1997) as the maximum likelihood estimate of a causal strength parameter. The value of causal support (Eq. 2) is a measure of whether a causal link exists. As Griffiths and Tenenbaum (2005) noted, "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" (p. 359).

Necessity and Sufficiency as Generic Priors

The second component of Bayesian inference, priors, will be especially important in guiding learning when data are sparse or noisy, as is often the case for naturalistic causal learning. In particular, the Bayesian formulation can take account of priors on the causal powers (i.e., w_0 and w_1). When learners have no obvious reason to have specific priors about weights (e.g., the power of a novel medicine to stop headaches), one might suppose that the priors are simply uniform (e.g., 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 variable. 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).

We propose that in the case of causal learning, people (and possibly other animals) have a prior favoring causes that are *necessary and sufficient* (e.g., a genetic defect on chromosome 4 is necessary and sufficient to cause Huntington's disease). The importance of necessity and sufficiency in causal inference was first discussed by J. S. Mill (1843). Causal necessity is the focus of the "but for" condition in law, and of the concept of attributable risk in epidemiology. In psychology, some have placed particular emphasis on sufficiency (e.g., Mandel & Lehman, 1998). Pearl (2000) reinterpreted various well-known causally-related measures in terms of probabilistic necessity and sufficiency (causal power as "probability of sufficiency"; attributable risk as "probability of necessity"; and ΔP as "probability of necessity and sufficiency"). Lien and Cheng (2000) proposed and provided evidence that a tacit goal of maximizing ΔP (i.e., necessity and sufficiency jointly), conditional on "no confounding", guides human induction of categories and causal powers at multiple hierarchical levels. However, previous researchers have not considered the possibility that the goal of maximizing the necessity and sufficiency of causes may provide relational generic priors that guide elemental causal induction.

Bayesian inference focuses on probabilistic rather than strictly deterministic relations. It would seem that most naturally-occurring causal relations are probabilistic, such that *C* is in fact neither necessary nor sufficient to produce E (e.g., the link between smoking and cancer). Nonetheless, a prior with weight peaks indicative of "approximately" necessary and sufficient causes (NS priors) would encourage causal networks that are inherently simple (ideally, one cause reliably predicts the effect). Such a prior would create a generic expectation in accord with what Holland, Holyoak, Nisbett and Thagard (1986, p. 160) termed "the "unusualness rule, unexpected events signal other unexpected events." For example, rats often show initial conditioning to a novel cue that precedes shock, even though the cue is in fact uncorrelated with shock (Rescorla, 1972). Readiness to "jump to causal conclusions" consistent with NS priors (assuming they can be overturned if contradicted by later experience) may have important survival value in a natural environment.

In the remainder of this paper we formulate the Bayesian model incorporating NS priors. We then summarize three human experiments, and compare model predictions using NS versus uniform priors with human causal judgments.

Bayesian Model with NS Priors

A Bayesian decision can be formalized to infer causal structure by assessing whether a causal relationship exists between *C* and *E* after observing contingency data *D*. The decision variable is obtained from the posterior probability ratio of Graphs 1 and 0 by applying Bayes' rule:

$$\log\frac{P(Graph1 \mid D)}{P(Graph0 \mid D)} = \log\frac{P(D \mid Graph1)}{P(D \mid Graph0)} + \log\frac{P(Graph1)}{P(Graph0)} \quad (1)$$

Griffiths and Tenenbaum (2005) defined the first term on the right of Eq. 1 (log likelihood ratio) as "causal support" (the second term, the log prior odds, is a constant). In general, support can be defined as the log posterior odds,

$$support = \log\frac{P(Graph1 \mid D)}{P(Graph0 \mid D)}, \quad (2)$$

a measure of the evidence that data *D* provide in favor of Graph 1 over Graph 0.

The likelihoods on graphs are computed by integrating out the unknown causal strengths w_0 and w_1, which are parameters in the range $\{0,1\}$ associated with the powers of *B* and *C*, respectively,

$$P(D \mid Graph1) = \int_0^1\int_0^1 P(D \mid w_0, w_1, Graph1)P(w_0, w_1 \mid Graph1)dw_0 dw_1$$

$$P(D \mid Graph0) = \int_0^1 P(D \mid w_0, Graph0)P(w_0 \mid Graph0)dw_0 \quad (3)$$

where $P(D \mid w_0, w_1, Graph1)$ and $P(D \mid w_0, Graph0)$ are the likelihood probabilities of the observed data given specified causal strengths and structures. $P(w_0, w_1 \mid Graph1)$ and $P(w_0 \mid Graph0)$ are prior probabilities that model the learners' beliefs about the values of causal strengths.

The likelihood terms are derived using the generating functions specified by the power PC theory. Let +/− indicate the value of the variable to be 1 vs. 0. For a Cheng model (noisy-OR) in which *B* and *C* are both potential generative causes, the probability of observing E is given by

$$P(e^+ \mid b, c; w_0, w_1) = 1 - (1 - w_0)^b (1 - w_1)^c \quad (4)$$

$b, c \in \{0,1\}$ varies with absence vs. presence of *C* (*b* is always 1). In the preventive case, *B* is assumed to be

potentially generative (following the "no background preventers" assumption of the power PC theory) and *C* is potentially preventive. The resulting noisy-AND-NOT generating model for preventive causes is

$$P(e^+ \mid b, c; w_0, w_1) = w_0{}^b (1 - w_1)^c \quad (5)$$

If data *D* is summarized by contingencies *N(e,c),* the number of cases for each combination of presence vs. absence of the effect and cause, then the likelihood given causal strengths (w_0, w_1) and structures (Graph 0,1) is

$$P(D \mid w_0, w_1, Graph1)$$
$$= \binom{N(c-)}{N(e+,c-)} w_0^{N(e+,c-)} (1 - w_0)^{N(e-,c-)}$$
$$\binom{N(c+)}{N(e+,c+)} [1 - (1 - w_0)(1 - w_1)]^{N(e+,c+)} [(1 - w_0)(1 - w_1)]^{N(e-,c+)}$$
$$P(D \mid w_0, Graph0)$$
$$= \binom{N(c-)}{N(e+,c-)} \binom{N(c+)}{N(e+,c+)} w_0^{N(e+,c-)+N(e+,c+)} (1 - w_0)^{N(e-,c-)+N(e-,c+)} \quad (6)$$

where $\binom{n}{k}$ denotes the number of ways of picking *k* unordered outcomes from *n* possibilities.

The second component in Eq. 3 is the prior on causal strength, $P(w_0, w_1 \mid Graph1)$ and $P(w_0 \mid Graph0)$. Griffiths and Tenenbaum (2005) assumed that the priors on weights w_0 and w_1 follow a uniform distribution. Our guiding hypothesis is that generic priors will favor necessary and sufficient causes. Accordingly, we set priors favoring NS generative causes, with the prior distribution peaks for w_0, w_1 at 0,1 (*C* is an NS cause) and 1,0 (*B* is). We use the exponential formulation

$$P(w_0, w_1 \mid Graph1) = \frac{1}{Z}\left[e^{-\alpha w_0} e^{-\alpha(1-w_1)} + e^{-\alpha(1-w_0)} e^{-\alpha w}\right] \quad (7)$$

where α is a parameter controlling how strongly necessary and sufficient causes are preferred, and Z is a normalizing term that ensures the sum of the prior probabilities equals 1. When α = 0, the prior follows a uniform distribution, indicating no preference to any values of causal strength. Griffiths and Tenenbaum's (2005) support model is thus derived as a special case. The present formulation provides an analytic calculation of support values.

$P(w_0 \mid Graph0)$ is obtained as the marginal of $P(w_0, w_1 \mid Graph1)$ by integrating out w_1,

$$P(w_0 \mid Graph0) = \frac{1}{Z}\left[e^{-\alpha w_0} + e^{-\alpha(1-w_0)}\right] \quad (8)$$

In the preventive case *B* is again assumed to be generative, hence only *C* could be a preventer (i.e., B and C do not compete). Evidence for *C* as an NS preventer will be clearest when *B* is a sufficient generative cause (w_0 = 1), yielding a likelihood peak for w_0, w_1 at 1,1:

$$P(w_0, w_1 \mid Graph1) = e^{-\alpha(1-w_0)} e^{-\alpha(1-w_1)} / Z \quad (9)$$

where α and Z are defined as in Eq. 7. As in the generative case, $P(w_0 \mid Graph0)$ is obtained as the marginal of $P(w_0, w_1 \mid Graph1)$:

$$P(w_0 \mid Graph0) = e^{-\alpha(1-w_0)} / Z \quad (10)$$

By substituting Eqs. 6 ~ 10 into Eq. 3, we can incorporate NS priors into computation of support for a causal link (Eq. 2). Fig. 2 depicts the prior distributions used in generative and preventive cases.

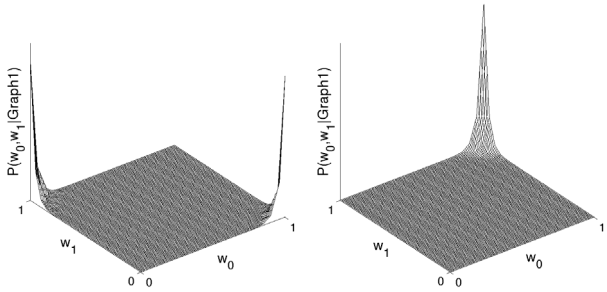

Figure 2: Prior distributions over w_0 and w_1 with NS priors. Left: Generative case, $\alpha = 30$ (peaks at 0,1 and 1,0); right: Preventive case, $\alpha = 30$ (peak at 1,1).

Overview of Experiments 1-3

Methods

Materials and procedure were very similar across all 3 experiments. Experiments 1-2 are from Liljeholm (2006). A simultaneous presentation format, adapted from that used by Buehner, Cheng and Clifford (2003, Ex. 2), was used to minimize memory demands and other processing issues extraneous to causal inference (see Fig. 3). The cover story always involved a set of allergy patients who either did or did not have a headache (*E*), and either had or had not received a new allergy medicine (*C*); the query concerned whether as a side effect the medicine caused headache (generative conditions) or relieved headache (preventive conditions). Each patient was represented by a cartoon face that was either frowning (headache) or smiling (no headache). The data were divided into 2 subsets, each an array of faces. The top subset represented patients who had not received the medicine; the bottom subset represented patients after receiving the medicine.

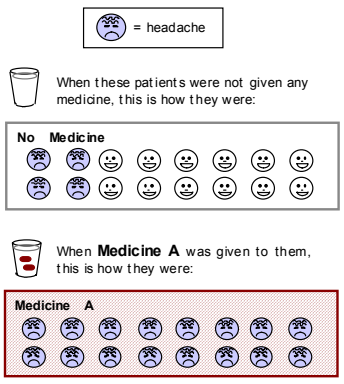

Figure 3. Example of an experimental display, showing patients who had not (top) or had (bottom) received an allergy medicine, and who either had or had not developed headaches.

The specific contingency conditions in each experiment are shown in Figs. 4-5. The code in the figures indicates number of patients with headache out of number who had not received the medicine (i.e., base rate of the effect), and number with headache out of number who did receive the medicine. The number of cases in the sample was varied. In the figures and all analyses, generative and preventive conditions are identical except that the frequencies of headache and no headache are transposed. For example, the generative case 2/8, 8/8, where $P(E|\tilde{C}) = .25$, $P(E|C) = 1$, power = 1, is matched to the symmetrical preventive case 6/8, 0/8, where $P(E|\tilde{C}) = .75$, $P(E|C) = 0$, power = 1. Ex. 1 included a series of contingency conditions in which the causal power of the medicine was 1 but the base rate of headache was varied, plus additional conditions with lower causal power.

The specific query regarding existence of a causal link varied across experiments. In Ex. 1 the query (generative conditions) was, "How likely is it that this medicine produces headaches?" with the response being a numerical rating on a line marked in units of 10 from 0 (extremely unlikely) to 100 (extremely likely). For preventive conditions, "produces" was replaced by relieves". The dependent measure was the rating in each condition. In Ex. 2 the query was, "Does this medicine cause headache? Rate how confident you are that this medicine causes headache" on a 100-point confidence scale. The dependent measure was the rating in each condition. In Ex. 3, 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).

Participants were UCLA undergraduates in the Psychology Department subject pool. Generative versus preventive conditions in Ex. 1 was a between-subject variable. In Ex. 1-2, contingency condition was a within-subjects variable, with order of presentation randomized. In Ex. 3 each participant evaluated a single condition. The data points for humans shown in Figs. 4-5 are each mean ratings based on responses from 20-33 participants.

Judgment Patterns

Before presenting the modeling results, it will help to characterize the major factors that influenced link judgments for both generative and preventive conditions (see Figs. 4-5). (1) *Causal power*: high power led to higher confidence there is a link. (2) *Sample size*: an overall larger sample tended to yield higher confidence (a surprisingly weak but statistically reliable factor in Ex. 1). (3) *Base rate of effect, $P(E+|C-)$*: confidence was higher when the base rate was more optimal for revealing any influence of the candidate cause, where the optimal base rate is 0 for the generative case and 1 for the preventive case. More optimal base rates lead to a larger "virtual sample" (Liljeholm, 2006), defined as the number of cases in which C could potentially reveal its influence; the complementary maximally suboptimal base rates lead to ceiling effects such that the power of C cannot be determined from the data. (4) *Direction of causation*: In Ex. 1, there was evidence of a possible interaction between causal direction and contingency condition. In particular, for conditions where $w_1 = 1$, preventive ratings tended to be higher than generative ratings when the base rate was far from optimal, with the difference diminishing as the base rate approached optimal. A comparison of the direction effect for the conditions in which the generative base rate was .75 (.25 preventive) vs. .25 (.75 preventive) yielded a significant interaction, $F(1, 51) = 4.71$, $p = .035$. Similar differences between preventive and generative judgments have been observed for causal strength judgments (Liljeholm, 2006; Wang & Fu, 2005).

Model Fits to Human Causal Judgments

Data from all 3 experiments were fit using the Bayesian model with either NS or uniform priors. An α value of 30 for NS priors was selected using data from Ex. 1, and then held constant in fitting data from Ex. 2-3. The model with uniform priors ($\alpha = 0$) is identical to that of Griffiths and Tenenbaum (2005). For both NS and uniform priors, support values were scaled to human data (a 100-point confidence scale) using a best-fitting power transformation (the same procedure employed by Griffiths & Tenenbaum).

Figs. 4-5 each show the data for human causal judgments (top) along with predictions based on NS priors (middle) and uniform priors (bottom). Ex. 1 tested 30 contingency conditions (15 generative and 15 preventive) with sample sizes of 32 (left side of Fig. 3) and 128 (right side). Although both Bayesian models fit the human data reasonably well, the overall correlation was substantially higher with NS priors ($r = .94$) than with uniform priors ($r = .71$).

Two qualitative aspects of the data favor the model with NS priors. First, NS priors capture the fact that human judgments of confidence in a causal link were more sensitive to causal power and $P(E+|C-)$ (base rate of the effect; e.g., increasingly optimal across left 6 contingencies in Fig. 4) than to sample size. Uniform priors place relatively greater weight on sample size. Second, NS priors capture the apparent asymmetry between generative and preventive judgments for cases matched on causal power and optimality of the base rate. For the human data, for 9 of the 10 matched conditions in which the base rate is non-optimal, the preventive rating exceeds the generative case. The asymmetric NS priors (1 peak for preventive causes, 2 for generative) capture this subtle interaction between preventive and generative judgments. In contrast, the model with uniform priors (like all previous formal models of causal judgments) predicts strict equality of matched generative and preventive conditions.

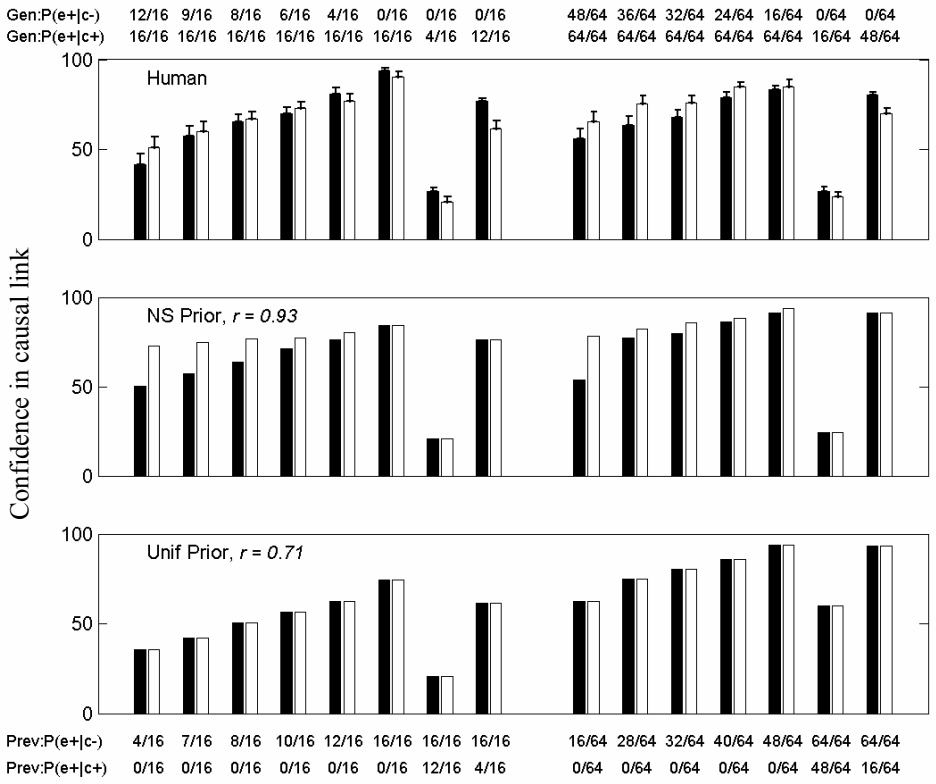

Figure 4: Confidence in a causal link (Ex. 1). Numbers along top show stimulus contingencies for generative cases; those along bottom show contingencies for matched negative cases. Top: Data from Ex. 1 (error bars indicate 1 standard error); middle: Predictions of Bayesian model with NS priors, $\alpha = 30$; bottom: Predictions with uniform priors, $\alpha = 0$.

Ex. 2 provided a further test of the relative potency of power and sample size as determinants of human causal judgments. This study employed two intermediate contingencies (powers of .4 and .67) at sample sizes 36 and 72 (generative conditions only). As shown in Fig. 5A, NS priors provided a far better fit to the human data (r = .97) than did uniform priors (r = .20). As in Ex. 1, NS priors capture the greater potency of power relative to sample size, whereas uniform priors erroneously predict the opposite trade-off.

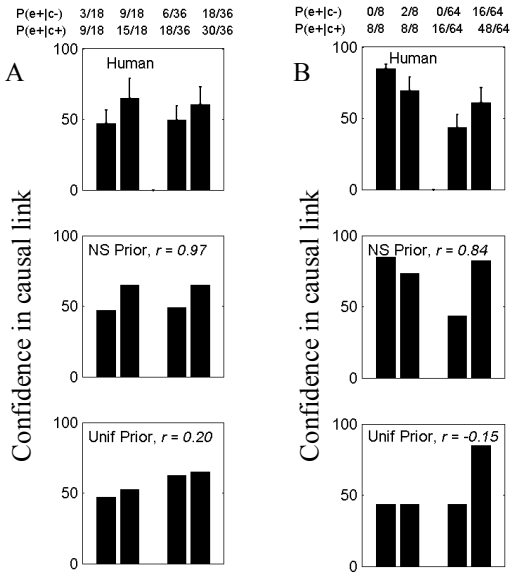

Figure 5. Confidence in a causal link. A: Ex. 2. B: Ex. 3. See Fig. 4 caption for additional information.

In the extreme, when the presented contingencies closely match the NS priors, the model with these generic priors predicts that people will be highly confident in the presence of a causal link after only a few observations. Ex. 3 was designed to test this prediction, comparing judgments for contingencies close to NS priors with a small sample size of 16 to contingencies far from NS priors with a substantially larger sample size of 128. As shown in Fig. 5B, NS priors again provided a much better fit (r = .84) than did uniform priors (r = –.15). As predicted, people placed much greater weight on match to NS priors than on sample size. In the most dramatic case, where the data fit the generative peak at w_0 = 0, w_1 = 1, human mean confidence was 85 on the 100-point scale after just 16 observations. NS priors closely match the human level of high confidence, whereas uniform priors erroneously predict a confidence level below 50. Moreover, uniform priors generate the wrong ordinal ranking of this favorable contingency relative to the rightmost condition in Fig. 5B (a case of lower power with a much high sample size).

Conclusions and Future Directions

We have established that a Bayesian formulation of causal inference that incorporates (1) a theory of learners' model of the generating model for binary causal variables and (2) generic priors favoring necessary and sufficient causes can explain the pattern of human causal judgments about existence of causal links. In contrast, a formulation assuming uniform priors (Griffiths & Tenenbaum, 2005) is unable to account for key findings. Humans place greater weight on match to NS priors than on size of the sample of

observations, and their causal judgments reveal a systematic interaction between preventive and generative ratings. NS priors are a special case of a general preference for simplicity in causal networks (cf. Novick & Cheng, 2004, p. 471).

The present Bayesian formulation, like that of Griffiths and Tenenbaum (2005), is based on a noisy-OR and noisy-AND-NOT generating model (Cheng model). Griffiths and Tenenbaum also discussed an alternative formulation based on a linear generating model that yields ΔP (i.e., $P(E+|C+) - P(E+|C-)$) as a strength measure. This model gives an incoherent account of independent causal influence (Cheng, 1997; Cheng et al., in press). It is clear the linear model will fail for the data modeled in the present paper. To take one simple example, each contingency in Ex. 2 (Fig. 5A) is equated for ΔP (.33); accordingly for paired conditions at each sample size, values of $P(E+|C-)$ and $P(E+|C+)$ vary symmetrically around .5. Since generative priors (either uniform or NS) for w_0 and w_1 are also symmetrical around .5, for these contingencies the linear model with either set of priors will necessarily predict support values that vary only with sample size. Clearly, however, people's confidence ratings varied with power within each sample-size condition even though ΔP was constant.

A major advantage of the Bayesian formulation of causal learning, when coupled with the concept of causal power, is that it is *compositional*: it allows the formulation of coherent answers to a wide variety of causal queries. Here we have focused on modeling support for a causal link, but the same formulation can also be used to model judgments of causal strength and confidence in strength judgments. Additional work will be required to extend the formulation to situations involving multiple candidate causes, potential interactive influences among causes, sequential presentation of data, and diagnostic inference from observed effects to possible causes.

Acknowledgments

Preparation of this paper was supported by a grant from the W. H. Keck Foundation to AY, NIH grant MH64810 to PC, and NSF grant SES-0350920 to KH.

References

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.

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

Cheng, P. W., Novick, L. R., Lijeholm, M., & Ford, C. (in press). Explaining four psychological asymmetries in causal reasoning: Implications of causal assumptions for coherence. In M. O'Rourke (Ed.), *Topics in contemporary philosophy (Vol. 4): Explanation and causation.* Cambridge, MA: MIT Press.

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

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

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

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

Liljeholm, M. (2006). Structure learning, parameter estimation and causal assumptions. Ph.D. dissertation, UCLA Department of Psychology.

Lu, H., & Yuille, A. (2006). Ideal observers for detecting motion: Correspondence noise. *Proceedings of Neural Information Processing Society*.

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

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.

Mill, J. S. (1843). *System of logic*, Vol. 1. London: John Parker.

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

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

Rescorla, R. A. (1972). Informational variables in Pavlovian conditioning. In G. H. Bower (Ed.), *The psychology of learning and motivation*. New York: Academic Press.

Shanks, D. R. (2004). Judging covariation and causation. In D. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making*. Oxford, UK: Blackwell.

Wang, M. Y., & Fu, X. L. (2005). Causal judgments in the trial-by-trial presentation. *Acta Psychologica Sinica*, *37*, 51-61.

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

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