Title

Bayesian Models of Judgments of Causal Strength: A Comparison

Permalink

https://escholarship.org/uc/item/7mx3b8rg

Authors

Lu, Hongjing
Yuille, Alan L.
Liljiholm, Mimi
et al.

Publication Date

2007-09-23

Peer reviewed

Bayesian Models of Judgments of Causal Strength: A Comparison

Hongjing Lu (hongjing@hkucc.hku.hk)
Department of Psychology, University of Hong Kong, Pokfulam Road, Hong Kong

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 formulate four alternative Bayesian models of causal strength judgments, and compare their predictions to two sets of human data. The models were derived by factorially varying the generating function for integrating multiple causes (based on either the power PC theory or the ΔP rule) and priors on strengths (favoring necessary and sufficient (NS) causes, or uniform). The models based on the power generating function provided much better fits than those based on the linear function. The models that included NS priors were able to account for subtle asymmetries between strength judgments for generative and preventive causes.

Keywords: Bayesian inference; causal strength; causal power

Strength Judgments in a Bayesian Framework

Humans and other animals are able to use contingency information to judge the potency of a cause that acts to produce or prevent an associated effect. For example, the HIV virus almost always leads to development of the disease AIDS (high causal strength), whereas smoking a pack of cigarette every day for a year leads to cancer with some small probability (low causal strength). In the simplest possible causal situation, a single binary candidate cause, C, varies along with a binary effect, E, against a constant background, B, that also might produce E (Figure 1). We assume B and C occur prior to E. The available contingency data takes the form of a tally of the frequencies with which E does or does not occur in the presence of either B alone (base rate of the effect) or B and C together. How can such contingency information be translated into an estimate of the causal strength of C?
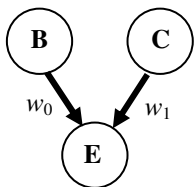

Figure 1. A simple causal graph. B, C, and E are binary variables. Weights w_0 and w_1 in Graph 1 indicate causal strength of the background cause, B, and the candidate cause, C, respectively.

Judging causal strength can be formalized as a Bayesian problem of parameter estimation within a fixed causal graph as shown in Figure 1 (Griffiths & Tenenbaum, 2005; Jaynes, 2003). Within the Bayesian framework, inference depends jointly on the *likelihoods* of data given alternative hypotheses, and on *priors* for these hypotheses. For strength judgments, likelihoods depend on the causal *generating function*, i.e., how do the influences of potential causes B and C in Graph 1 combine to influence E? The relevant priors are initial probabilities assigned to possible values of the weights w_o and w_1 (representing causal strengths) on the causal links for B and C, respectively.

Derivation of Bayesian Models

Based on observation of contingency data D, a Bayesian model can be formalized to infer causal strength. The Bayesian model is able to assess the probability distribution of w_1 so as to quantify statistical properties of the causal strength of candidate cause C to produce or prevent E. In this paper we compare the average human strength rating with the mean of w_1 in the causal structure shown in Figure 1. The mean of w_1 is determined by

$$\overline{w_1} = \int_0^1 w_1 P(w_1 \mid D)\cdot \tag{1}$$

The posterior distribution $P(w_1 \mid D)$ is obtained by applying Bayes rule,

$$P(w_1 \mid D) = \int_0^1 P(w_0, w_1 \mid D) dw_0 = \int_0^1 \frac{P(D \mid w_0, w_1) P(w_0, w_1)}{P(D)} dw_0, \tag{2}$$

where $P(D \mid w_0, w_1)$ is the likelihood term. $P(w_0, w_1)$ corresponds to 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. Let $+/-$ indicate the value of the variable to be 1 (present) versus 0 (absent). The likelihood term $P(D \mid w_0, w_1)$ is given by

$$P(D \mid w_0, w_1) = \binom{N(c^-)}{N(e^+, c^-)} \binom{N(c^+)}{N(e^+, c^+)} \prod_{e,c} P(e \mid b, c; w_0, w_1)^{N(e,c)} \quad (3)$$

where $b, c, e \in \{0,1\}$ denotes the absence and the presence of the causes B, C, and the effect E. $\binom{n}{k}$ denotes the number of ways of picking k unordered outcomes from n possibilities.

Alternative Generating Functions Griffiths and Tenenbaum (2005) pointed out that Bayesian models of causal judgments can be constructed using either of two generating functions derived from models in the psychological literature. The generating function adopted by Griffiths and Tenenbaum in their "causal support" model is the *power* generating function, derived from the power PC theory (Cheng, 1997; see Glymour, 2001). 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,

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

It is assumed that $b = 1$ because the background cause B is always present in the experimental setup. In the preventive case, B is again assumed to be potentially generative (following the power PC theory, which specifies that the background must not include preventive causes), whereas C is potentially preventive. The resulting noisy-AND-NOT generating function for preventive causes is

$$P(e^+ \mid b, c; w_0, w_1) = w_0 b - w_0 w_1 bc. \quad (5)$$

For convenience we will refer to Eqs. 4-5 together as the power generating function. 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].

Using the power generating function, Cheng (1997) derived quantitative predictions for judgments of causal strength. Let q represent a point estimate of the value of causal *power*. When certain assumptions are satisfied, the predicted value of causal power for a generative cause is

$$q_G = \frac{\Delta P}{1 - P(e^+ \mid c^-)}, \quad (6)$$

and the predicted value of power for a preventive cause is

$$q_P = \frac{-\Delta P}{P(e^+ \mid c^-)}, \quad (7)$$

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

$$\Delta P = P(e^+ \mid c^+) - P(e^+ \mid c^-). \quad (8)$$

Griffiths and Tenenbaum (2005) showed that causal power (q in Eqs. 6-7) corresponds to the maximum likelihood estimate for the random variable w_1 on a fixed graph (as shown in Figure 1) under the power generating function.

The term $P(e^+ \mid c^-)$ in the denominator of Eqs. 6-7 is often termed the *base rate of the effect*, as it gives the prevalence of the effect under background conditions in the absence of the candidate cause. The base rate determines the value of w_0 in the causal structural graph shown in Figure 1. Note that the derivation of causal power for generative causes (Eqs. 4, 6) versus preventive causes (Eqs. 5, 7) is inherently asymmetrical with respect to the base rate of the effect.

An alternative generating function can be derived directly from ΔP (Eq. 8), which has been interpreted by some theorists as an estimate of causal strength (Jenkins & Ward, 1965). Under certain conditions, when learning is at asymptote the ΔP rule is equivalent to the Rescorla-Wagner associative learning model (see Danks, 2003), which has been advanced as a model of causal inference (Shanks & Dickinson, 1987). The ΔP model yields a linear generating function,

$$P(e^+ \mid b, c; w_0, w_1) = w_0 b + w_1 c \quad (9)$$

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. Eq. 9 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 Eq. 9 yields ΔP as the maximum likelihood estimate of w_1 when substituted for Eqs. 4-5.

Alternative Priors The second component in Eq. 2 is the prior on causal strength, $P(w_0, w_1)$, within the causal structure in Figure 1. 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 and Tenenbaum (2005) adopted uniform priors in their causal support model.

An alternative proposal is that people have priors for *necessary and sufficient* (NS) causes (Lu et al., 2006). Our *NS power* model integrates the power generating function with *generic* priors (cf. Lu & Yuille, 2006) about the relationship between the powers of alternative potential causes. We make the assumption that people prefer causal networks that are relatively simple (Novick & Cheng, 2004, p. 471) and that people have a deterministic bias regarding causal strength. Causal simplicity (Chater & Vitányi, 2003) potentially manifests itself in multiple ways, which likely include a preference for fewer causes (Lombrozo, 2007) and for causes that do not involve interactions (Novick & Cheng, 2004; Liljeholm & Cheng, in press). Deterministic causal preference biases causal strength towards 0 and/or 1. NS priors imply that people have a prior belief favoring causes that are *necessary and sufficient* (e.g., a genetic defect on chromosome 4 is necessary and sufficient to cause Huntington's disease). But rather than being a strict logical condition, NS priors are assumed to be probabilistic. Pearl (2000) interpreted generative causal power (Eq. 3) as "probability of sufficiency," and ΔP (Eq. 6) as "probability of necessity and sufficiency." (For preventive causes the analogous quantities are preventive causal power and $-\Delta P$, respectively.) Developmental data provide support for the

assumption of NS priors. Recent evidence indicates that preschool children tacitly believe in "causal determinism", inferring unobserved causes to explain apparently stochastic patterns of effects (Schultz & Sommerville, 2006).

For the generative case, the background B and candidate C are both potentially generative, and hence will implicitly compete as alternative NS 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) (see Figure 2A). We specify the priors using a mixed distribution with exponential functions,

$$P(w_0, w_1 \mid gen) = \frac{e^{-\alpha w_0 - \alpha(1-w_1)} + e^{-\alpha(1-w_0) - \alpha w_1}}{Z(\alpha)}, \quad (10)$$

where α is a parameter controlling how strongly necessary and sufficient causes are preferred. When $\alpha = 0$, the prior follows a uniform distribution, indicating no preference to any values of causal strength. $Z(\alpha) = 2(1 - e^{-\alpha})^2 / \alpha^2$ denotes a normalization term that ensures the sum of the prior probabilities equals 1. Figure 2A depicts the shape of a distribution when $\alpha = 5$.
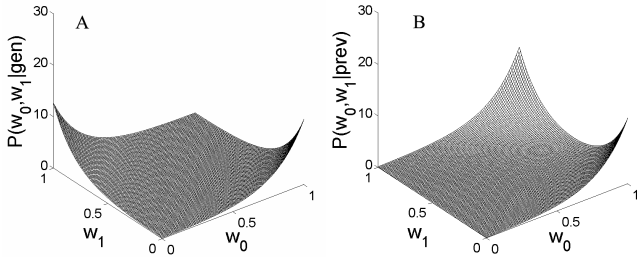

Figure 2. Prior distributions over w_0 and w_1 with NS priors. A: Generative case, $\alpha = 5$ (peaks at 0,1 and 1,0); B: Preventive case, $\alpha = 5$ (peaks at 1,1 and 1,0).

The NS prior will differ for the preventive case (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 NS causes in the preventive case. The issue of prevention will arise under the assumption that the effect is being generated; hence the peak weight of w_0 for the background cause B (the only possible generative cause) is assumed to be biased towards 1. The maximum probability of necessity and sufficiency for C as a preventer will then obtain when B is a sufficient generative cause, $w_0 = 1$, yielding a distribution peak for w_0, w_1 at 1,1. If C is not sufficient, the alternative consistent with causal determinism is that it is completely ineffective, yielding an additional peak at 1,0. We again use an exponential formulation,

$$P(w_0, w_1 \mid prev) = \frac{e^{-\alpha(1-w_0) - \alpha(1-w_1)} + e^{-\alpha(1-w_0) - \alpha w_1}}{Z(\alpha)}, \quad (11)$$

where all parameters are defined as in Eq. 10.

Note that the two peaks of the NS priors for the preventive case (Figure 2B) are not symmetrical with those for the generative case (Figure 2A). As we will see, the asymmetrical NS priors for generative versus preventive causes yield systematic asymmetries in human causal judgments as a function of causal direction.

In summary, the factorial combination of two alternative generating functions (power versus linear) and two alternative priors (NS or uniform) yields four alternative Bayesian models: Model I (power, NS), Model II (power, uniform), Model III (linear, NS), and Model IV (linear, uniform). Model I corresponds to the NS power model (Lu et al., 2006) when adapted to estimate causal strength. Model II corresponds to the causal support model (Griffiths & Tenenbaum, 2005) when adapted to estimate 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). Model III, identical to Model IV except with NS priors, has never been previously considered.

Simulations of Human Strength Judgments

We tested these four models by comparing the predictions of each for two data sets of human judgments of causal strength. Methodological issues arise in selecting data for quantitative modeling of strength judgments. Many studies have used rating scales to assess causal strength. As pointed out by Buehner et al. (2003), such scales may be ambiguous, leading participants to give responses that conflate causal strength with reliability. 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). The two data sets we selected for modeling used this type of query, coupled with summary displays of contingency information in which individual cases are presented in a single organized display (see Figure 3 for an example). Such presentations provide a vivid display of individual cases, making salient the frequencies of the various types of cases, while minimizing memory demands.

Data Set 1: "Headache" Cover Story

We first modeled a large data set from a study by Liljeholm and Cheng (2007, Experiment 1).

Method Fifty-two undergraduates at the University of California, Los Angeles (UCLA) were assigned in equal numbers to generative and preventive conditions. 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 "remove" 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, might 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 3, 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 before receiving the mineral and depicted $P(e+|c-)$; the bottom subset represented patients after receiving the mineral and depicted $P(e+|c+)$.
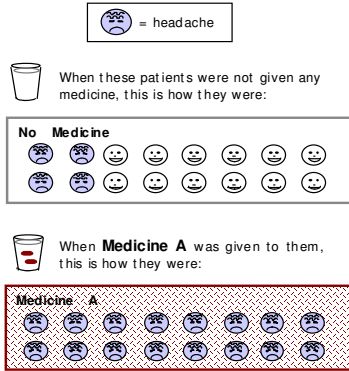
Figure 3. Example of a "headache" display, showing patients who had not (top) or had (bottom) received a mineral used in an allergy medicine, and who had or had not developed headaches.

Contingency conditions were varied within-subjects. Two samples sizes (16 and 64) were combined with two causal powers, .25 and 1, and three base rates: 0 (for generative; 1 for preventive), .25, and .75, yielding a total of 24 conditions (see Figure 4). The code in Figure 4 indicates number of patients with headache out of total number before receiving the mineral (i.e., base rate of the effect), and number with headache out of total number after receiving the mineral (where the mineral is *C* and headache is *E*). In Figure 4, generative and preventive conditions are identical except that the frequencies of headache and no headache 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.

Before answering the strength query, participants were asked if "The mineral **has absolutely no influence** on headache." Strength ratings were not obtained for those participants who agreed with this assertion. The subsequent query (generative conditions) was, "Suppose that there are 100 people that **do not** have headaches. If this mineral was given to these 100 people, how many of them would have headaches?" The preventive version simply substituted "**do**" for "**do not**" and "have" for "not have". Participants had been instructed not to provide any numerical rating when selecting the first answer option, as well as to not put a zero rating when selecting the second answer option. The dependent measure of causal strength was the average of numerical rating (1-100) elicited in each condition for the second query.

Fits of Bayesian Models
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). 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 and Cheng (2007), we assume that mean strength ratings on the 100-point scale approximate a ratio scale of causal power.[1]
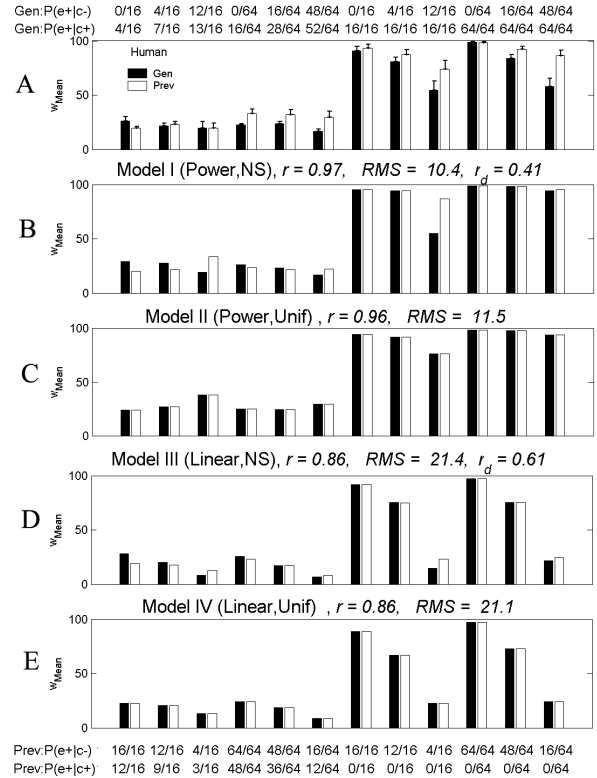
Figure 4. Strength ratings (Data Set 1). Numbers along top show stimulus contingencies for generative cases; those along bottom show contingencies for matched negative cases. A: Mean human ratings (error bars indicate 1 standard error); B~E: Predictions from four models.

Hence, a successful model must aim to account for the actual values obtained for human strength judgments, without any further data transformation. Accordingly, we report model fits not only based on correlations, but also on root-mean-squared (RMS) deviations from the human data. In addition, the models with NS priors predict systematic interactions 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

---

[1] The assumption of a ratio scale is likely to break down for strength estimates near the extremes (0 or 100 on the scale) due to measurement issues (errors necessarily fall one side of the true mean).

predicted difference score is always 0 for Models II and IV, which assume uniform priors; hence r_d is not computable.

The human data (Figure 4A) were well fit overall by Models I and II based on the power generating function, either with NS priors (Figure 4B) or uniform priors (Figure 4C; $r = .97$ and .96, respectively). Model I (NS power) has a slight advantage in terms of lower RMS, and in addition yields a positive correlation with the difference in strength ratings for matched generative and preventive contingencies ($r_d = .41$). Models III and IV based on the linear generating function (Figure 4D-E) yielded substantially poorer overall fits ($r = .86$ for each), roughly doubling the RMS relative to the models based on the power generating function. The reason for the poor fits of the linear models is that they erroneously predict that human strength judgments will asymptote at values corresponding to values of ΔP, whereas human judgments actually approach values of causal power at asymptote. The linear Model III with NS priors does, however, yield a positive correlation with difference scores for generative versus preventive causes ($r_d = .61$).

Data Set 2: "DNA" Cover Story

For generality, we performed an experiment to obtain strength ratings using a different cover story.

Method Seventy-four UCLA undergraduates served in the study. The cover story concerned a bio-genetics company testing the influence of various proteins on the expression of a gene. Participant 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" of the same basic sort used the "headache" study (see Figure 5). One sample of DNA strands had not been exposed to a particular protein, while the other sample of DNA strands had been exposed to that protein. The 16 contingencies used in the experiment are shown in Figure 6. Causal direction was varied between-subjects, contingency within-subjects.

Strength judgments were obtained from all participants. 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?" The preventive query was identical except that "OFF" was replaced by "ON" and "TURNED ON" by "TURNED OFF".

Fits of Bayesian Models

As for the data for the "headache" cover story, the human data based on the "DNA" cover story (Figure 6A) was well fit overall by Models I and II based on the power generating function, either with NS (Figure 6B) or uniform priors (Figure 6C; $r = .98$ for each). The RMS was very low for

both models, with a slight advantage (less than 1 point on the 100-point scale) for Model II. However, Model I (NS power) yielded a substantial positive correlation with the difference in strength ratings for matched generative and preventive contingencies ($r_d = .80$), whereas Model II with
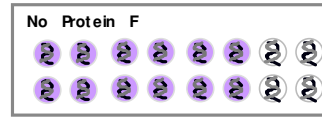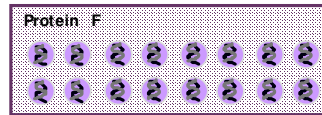

Figure 5. Example of a "DNA" display, with DNA strands that had not (top) or had (bottom) been exposed to a protein, and indicating whether a gene was off or on.
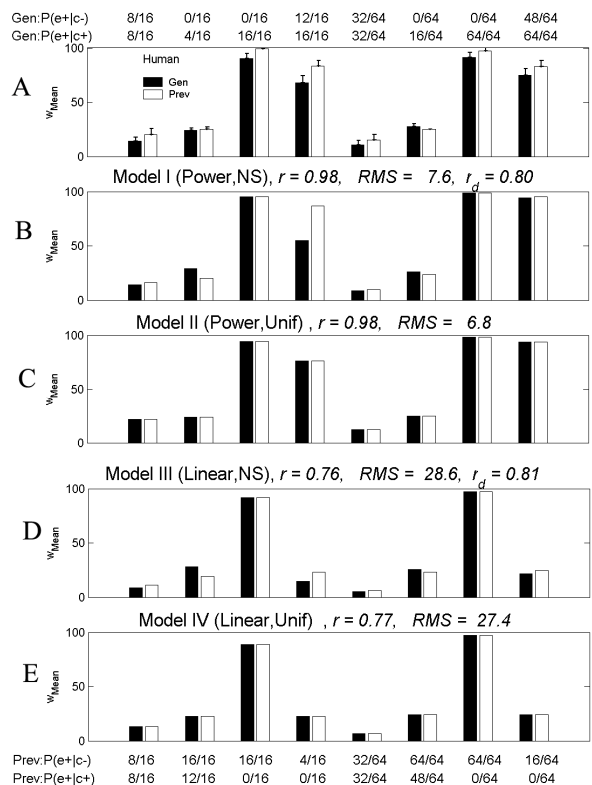

Figure 6. Strength ratings (Data Set 2). Same conventions as in Figure 4.

uniform priors is completely unable to account for the pattern of interactions with causal direction.

Once again, Models III and IV based on the linear generating function (Figure 6D for NS priors, Figure 6E for uniform priors) yielded substantially poorer overall fits ($r = .77$ and .76, respectively), roughly quadrupling the RMS relative to the models based on the power generating

function. Model III with NS priors did, however, yield a positive correlation with difference scores for generative versus preventive causes ($r_d = .81$).

General Discussion

In summary, the best overall account of human strength judgments for both the "headache" and "DNA" data sets was provided by the NS power model (Model I), which combines the power generating function with NS priors. NS priors provide the only formal account to date of asymmetries between causal judgments for generative and preventive causes. Similar asymmetries have been observed for judgments of whether or not a causal link is present (structure judgments; Lu et al., 2006).[2]

The quantitative failure of the linear generating function (Models III and IV) confirms the negative conclusion that has been reached on the basis of more qualitative 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 directly in terms of ΔP (Jenkins & Ward, 1965) or indirectly in the Rescorla-Wagner model (Shanks & Dickinson, 1987).

An important meta-point is that there may be many potential "rational" models of a given cognitive task. The Bayesian framework simply derives rational predictions from stated theoretical premises: *if* a reasoner has certain entering causal beliefs, *and* believes that causes follow a certain function in generating their effects, *then* some pattern of rational causal judgments follows. The four Bayesian models we have considered here differ in their underlying theoretical premises, and hence in their empirical predictions. The Bayesian framework provides a natural formalism for deriving and comparing the quantitative predictions of alternative "rational" models.

Acknowledgments

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.

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.

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

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.

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.

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

Liljeholm, M., & Cheng, P. W. (in press). When is a cause the "same"? Coherent generalization across contexts. *Psychological Science*.

Liljeholm, M., & Cheng, P. W. (2007). The influence of actual and virtual sample size on confidence and causal strength judgments. UCLA Department of Psychology.

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

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

Lu, H., Yuille, A., 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.

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

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.

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

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.

---

[2] NS priors would appear to be weaker for judgments of strength ($\alpha = 5$) than for structure ($\alpha = 30$; see Lu et al., 2006). We have since reformulated our structure model to hold NS priors constant across structure and strength judgments, while adding an additional component of priors specific to structure judgments.