Estimating human priors on causal strength

Saiwing Yeung (saiwing@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA. 94720-1650 USA

Abstract

Bayesian models of human causal induction rely on assumptions about people's priors that have not been extensively tested. We empirically estimated human priors on the strength of causal relationships using iterated learning, an experimental method where people make inferences from data generated based on their own responses in previous trials. This method produced a prior on causal strength that was quite different from priors previously proposed in the literature on causal induction. The predictions of Bayesian models using different priors were then compared against human judgments of strength of causal relationships. The empirical priors estimated via iterated learning resulted in the best predictions.
Keywords: Causal learning; Bayesian inference; Probabilistic judgment; Iterated learning

Introduction

Causal induction involves inferring the relationship between causes and effects. This problem has attracted the attention of cognitive scientists because it is an important skill that people rely on every day in order to understand the causal relationships in their environment. Traditionally, psychological models of human causal induction have focused on various schemes for comparing the probability of an effect occurring in the presence and absence of a cause (e.g., Ward & Jenkins, 1965; Cheng, 1997). However, recent work has explored connections between ideas from Bayesian statistics and human cognition, using causal graphical models to precisely define the problem of causal induction (Griffiths & Tenenbaum, 2005; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008) and to formalize the effects of prior knowledge (Griffiths & Tenenbaum, 2009). A key part of these Bayesian models is to precisely specify the prior knowledge that people have about the strength of causal relationships. In previous models of human causal induction, priors on the strength of causal relationships were specified in two ways — either as uniform priors by appealing to the principle of indifference (Griffiths & Tenenbaum, 2005), or as generic priors based on assumptions about the abstract properties of the causal system (Lu et al., 2008). In this paper, we present a new approach to estimating human priors on causal strength, using the method of *iterated learning* (Kalish, Griffiths, & Lewandowsky, 2007; Griffiths, Christian, & Kalish, 2008).

Iterated learning was originally proposed as a simple model of the cultural transmission of languages (Kirby, 2001). In this case, we imagine a chain of agents, where each agent observes data generated by the previous agent (such as a set of utterances), forms a hypothesis about the process that generated those data (such as a language), and then generates new data to pass to the next agent. If the agents select hypotheses using Bayesian inference, then as the chain gets longer the probability that an agent selects a particular hypothesis converges to the prior probability of that hypothesis (Griffiths & Kalish, 2007). Simulating this process of iterated learning in the laboratory thus provides a way to estimate people's priors (Kalish et al., 2007). In fact, there is no need for data to be passed between people — we can just generate the data that people see on one trial based on their responses in a previous trial (Griffiths et al., 2008).

The plan for the rest of the paper is as follows. The next section summarizes previous work on modeling human causal induction, focusing on analyses based on causal graphical models. We then introduce the basic ideas behind iterated learning and present our experimental investigation of human priors on causal strength. Next we compare the predictions produced by a model using the empirical priors with previous models. Finally we conclude the paper by discussing the implications of these results for understanding causal induction.

Models of human causal induction

The British philosopher David Hume pointed out that people are not "able to comprehend any force or power by which the cause operates, or any connexion between it and its supposed effect" (Hume, 1739/2004, p. 47), suggesting that causal relationships need to be inferred from the observed contingencies of cause and effect. A number of models have been proposed to account for how this inference might be made, with the goal of predicting human judgments about causal relationships from contingency data.

Models based on cause-effect probabilities

The ΔP model (Ward & Jenkins, 1965) proposed that human make inferences about causal strength based on the contrast between $P(e^+|c^+)$ and $P(e^+|c^-)$, where e and c represent the effect (or outcome) and cause, and superscripts of $^+$ and $^-$ represent their presence or absence. ΔP is formally expressed as $\Delta P = P(e^+|c^+)$ - $P(e^+|c^-)$. It captures the intuition that a cause is strong if it significantly increases the probability of the outcome occurring relative to its base rate.

Cheng (1997) argued that ΔP was just a measure of covariation and not one of causality. She further proposed the theory of causal power, in which human judgments of causal strength equals the probability of the cause in question produces the effect in the absence of all other causes. For example, the power model for a generative cause can be expressed as power $= \frac{\Delta P}{1-P(e^+|c^-)}$. The causal power model provided better fit than ΔP for some human data. However, there have been debates about the lack of fit to human data for both models (see Buehner & Cheng, 1997; Lober & Shanks, 2000).

Models based on Bayesian statistics

Griffiths and Tenenbaum (2005) proposed a Bayesian framework for studying causal induction. This framework uses causal graphical models to distinguish between *causal structure* – whether or not a link between two variables exists – and *causal strength* – the strength of that relationship. Griffiths and Tenenbaum gave a Bayesian account of learning causal structure. Lu et al. (2008) recently showed how this approach could be extended to infer causal strength.

Causal graphical models are probabilistic models in which a graph is used to denote the causal relationships between variables (Pearl, 2000; Spirtes, Glymour, & Scheines, 2001). In the graph, nodes represent variables and edges represent the causal connection between those variables. Following Griffiths and Tenenbaum (2005) and Lu et al. (2008), we focus on causal systems in which there are three variables: the background cause B, the potential cause C, and the effect E. Assuming both B and C can cause E, this relationship can be expressed in a graph in which there are edges going from both B and C to E. We assume B is always present and is generative, increasing the probability of the outcome, while C can be present or absent, and generative or preventive. We further assume that E cannot occur unless B or C caused it. Inferences are based on contingency data, which can be summarized in a 2×2 contingency table indicating the frequencies with which all combinations of the presence and absence of C and E co-occur.

Although the graph structure specifies the causal relationships among variables, the exact nature of those relationships is not clear without specifying their functional form. Noisy-OR (for generative causes) and noisy-AND-NOT (for preventive causes) parameterizations have been used to characterize the functional forms of causal relationships in previous models of causal induction (Cheng, 1997; Griffiths & Tenenbaum, 2005). Each cause is assumed to have the power to cause (or prevent) the effect independently, doing so with a probability that reflects its strength. We denote the strength of B and C as w_0 and w_1 respectively. The noisy-OR gives the probability of observing the effect E as $P(e^+|c;w_0,w_1) = 1-(1-w_0)(1-w_1)^c$ where c is a binary value representing the presence or absence of C, while the noisy-AND-NOT gives $P(e^+|c;w_0,w_1) = w_0(1-w_1)^c$.

Having specified the full causal model, we can use this model to infer the strength of B and C from the observed contingency data.[1] These data indicate the frequency with which cause and effect co-occur. We will use $N(e,c)$ to denote the number of cases falling into each cell of the contingency table, with e ranging over e^+ and e^-, and c ranging over c^+ or c^-. For any particular value of w_0 and w_1, the probability of the observed contingency data D is

$$P(D|w_0,w_1) = \prod_{e,c} P(e|c;w_0,w_1)^{N(e,c)} \quad (1)$$

[1] We focus on the problem of estimating causal strength, but the prior we estimate on causal strengths can also be used to infer causal structure (as in Griffiths & Tenenbaum, 2005; Lu et al., 2008).

where $P(e|c;w_0,w_1)$ is given by the noisy-OR or noisy-AND-NOT as above. We can thus compute a posterior distribution over w_0 and w_1 given D by applying Bayes' rule, with

$$P(w_0,w_1|D) \propto P(D|w_0,w_1)P(w_0,w_1) \quad (2)$$

where $P(w_0,w_1)$ is the prior on w_0 and w_1. Estimates of w_0 and w_1 can then be obtained by taking the posterior expectation, with $\bar{w}_i = \int_0^1 w_i P(w_0,w_1|D)\,\mathrm{d}w_i$ for $i \in 0,1$.

Priors on causal strength

In order to evaluate the posterior distribution in Equation 2, or to integrate over causal strength to evaluate causal structures as in Griffiths and Tenenbaum (2005), the prior on the causal strengths w_0 and w_1 needs to be specified. Griffiths and Tenenbaum assumed a uniform prior on both variables in their Bayesian structure learning model. However, Lu et al. (2008) argued that human reasoning are better approximated using a model that incorporated generic priors — a theoretically driven prior that makes systematic assumptions about the abstract properties of a system. They argued that people have preference for causal models with fewer causes (Lombrozo, 2007), and for causes that minimize complex interactions (Novick & Cheng, 2004). Based on these arguments, they specified the sparse and strong (SS) prior as $P(w_0,w_1) \propto e^{-\alpha(1+w_0-w_1)} + e^{-\alpha(1-w_0+w_1)}$ in the generative case and $P(w_0,w_1) \propto e^{-\alpha(1-w_0+1-w_1)} + e^{-\alpha(1-w_0+w_1)}$ in the preventive case (Lu et al., 2008). α in the formulae is a free parameter and is fixed at 5 in their analysis. In the generative case, this formulation gives higher prior probability when one of the causes (B or C) is very strong and the other is very weak; in the preventive case, this formulation gives higher prior probability when B is very strong and C is either very strong or very weak. Although Lu et al.'s model based on generic priors provided a good fit to human judgments in their experiments, there are infinitely many possible priors on causal strength, of which this is just a single possibility.

Using iterated learning to estimate human priors on causal strength

Previous work such as Lu et al.'s evaluated different proposals about priors on causal strength by testing predictions of specific models implementing those priors. Here, we take a different approach, using an experimental method based on iterated learning to directly estimate human priors on causal strength. When used as a model of cultural transmission, iterated learning refers to a process in which a sequence of agents each learns from data generated by the previous agent. Formally, the n-th agent observes data $d^{(n)}$ and forms a hypothesis $h^{(n)}$ about the process that generated those data, then goes on to generate the data $d^{(n+1)}$ which is given to the next agent. This defines a Markov chain on hypothesis-data pairs. If the agents select a hypothesis by sampling from the posterior distribution $p(h|d) \propto p(d|h)p(h)$ and then generate data by sampling from the corresponding likelihood function $p(d|h)$, then this Markov chain is a Gibbs sampler for the joint distribution

$p(d,h) = p(d|h)p(h)$. As n becomes large, the distribution of (d_n, h_n) converges to this joint distribution, and the probability that the nth learner selects a particular hypothesis h converges to $p(h)$ (for details, see Griffiths & Kalish, 2007).

Convergence of iterated learning to the prior suggests that it might be used as a method for empirically estimating human priors. Laboratory simulations of iterated learning, with data being passed between people, support this idea: Functions (Kalish et al., 2007) and concepts (Griffiths et al., 2008) transmitted through iterated learning quickly converge to forms that are consistent with priors established in previous research. However, there is no need for data to be transmitted between people for this to occur: A feedback process can be established for a single individual that has the appropriate statistical structure, where people form hypotheses about data that are generated based on their responses on previous trials. This kind of within-subjects design has previously been used to explore people's inductive biases in concept learning, producing equivalent results to a between-subjects design (Griffiths et al., 2008). We now explore whether a similar approach can identify people's prior on causal strength.

Methods

Participants. Participants were recruited from the University of California, Berkeley, subject pool, and online (Amazon Mechanical Turk). Participants from the subject pool received course credit, while online participants received a small payment. Only data from participants who completed at least 95% of the trials are included in the analysis. In the *generative* condition, there were 20 and 52 participants from the university subject pool and online respectively. In the *preventive* condition, there were 33 and 51 participants from these groups.

Stimuli and Procedure. Following Lu et al. (2008), we presented the experiment using a cover story of a bio-genetics company testing the influence of proteins on the expression of genes. The experiment was run in a web browser. In the *generative* condition, participants were told:

> In this experiment, please imagine that you are a researcher working for a bio-technology company and you are studying the relationship between genes and proteins concerning gene expression.
>
> Gene expression is the process by which information from a gene is used in synthesizing RNA or other proteins; it controls the structure and functions of cells or other genes. This process may or may not be modulated by the presence of proteins. You are given a number of gene/protein pairs and your job is to make assessments concerning the effect of these proteins on the expression of the genes.
>
> There are a number of trials in this experiment. Each trial involves a different gene and a different protein. In each trial, DNA strands extracted from hair samples would be exposed to the protein and the expression of the gene would be assessed. You will see the results from two samples of DNA strands. One sample consists of DNA strands that had not been exposed to the protein while the other sample consists of DNA strands that had been exposed to that protein. In both samples, you will see the number of gene expression resulted but no other information will be provided.

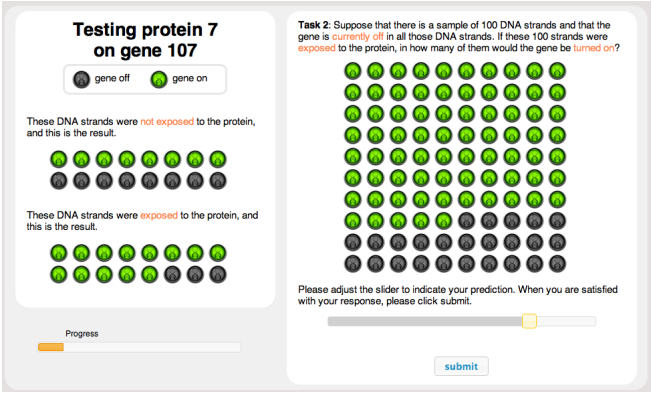

Figure 1: Screenshot of the *generative* condition of the experiment. The participant is assessing the strength of the cause C, w_1.

Participants then received instructions familiarizing them with the controls that they would use in the experiment. Each trial (50 total) was presented on a separate screen (see Figure 1). In each trial participants saw data in the form of two samples, one that was not exposed to the protein (c^-) and one that was (c^+). The data were presented graphically using pictures that showed the total number of DNA strands in each sample as well as the number that expressed the gene, providing complete contingency data $N(e,c)$.

After observing these contingencies, participants were asked to make two judgments involving hypothetical samples. The instructions were:

> Suppose that there is a sample of 100 DNA strands and these strands were not exposed to the protein, in how many of them would the gene be turned on?
>
> Suppose that there is a sample of 100 DNA strands and that the gene is currently 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?

These questions were phrased to elicit judgments of w_0 and w_1, based on stimuli from previous research (e.g., Lu et al., 2008). Participants responded using a slider. Live feedback showing the proportion of expressed genes was shown as the slider was moved. Participants could adjust the slider until they were satisfied with their response, before clicking the submit button to record their response and go to the next trial. The instructions for the *preventive* condition were similar, except that the protein was characterized as having preventive power, and the hypothetical sample used in the second question was assumed to have the gene currently expressed in all 100 DNA strands.

A within-subjects iterated learning design was used. There were four transmission chains with ten iterations each. The data that initiated each chain were generated by sampling a contingency table from predetermined initial values of w_0 and w_1. These initial values were chosen to be distinct so that the differences between responses from different chains could be used to diagnose convergence. The initial data for the four dependent chains were drawn from distributions with (w_0, w_1)

parameters of $(0.2,0.2)$, $(0.2,0.8)$, $(0.8,0.2)$, and $(0.8,0.8)$. Each contingency table had a total of 16 cases where the cause was present, and 16 where the cause was absent, with the number of times the effect occurred being generated via a binomial draw with parameter $P(e^+|c;w_0,w_1)$. In all subsequent trials, data were generated based on the w_0 and w_1 values the participant produced in the previous trial in the same chain. These values were taken directly from the estimates that the participants produced in response to the two questions about hypothetical samples. For example, if on iteration n of a particular chain the participant's responses were f_0 and f_1 (out of 100) for the two questions, then the data presented at iteration $n+1$ of chain would be drawn using $w_0 = f_0/100$ and $w_1 = f_1/100$.

To evaluate the performance of different Bayesian models, we added a fifth chain in which the contingencies that were presented did not depend on participants' previous responses. This is not really a "chain" as all of these trials were in fact independent, but we retain the name for convenience in exposition and in contrast to the dependent chains. This chain provided the additional benefit of preventing participants from being able to easily guess or approximate the stimulus generation algorithm (Griffiths et al., 2008). For the independent chain both w_0 and w_1 were sampled from a uniform distribution on $(0,1)$ on each trial, and were then used to generate contingencies as in the dependent chains. There were thus a total of five chains and 50 trials per participant. The order of trials between chains was randomized within each iteration.

Results

The results from the university and online subject pools were similar to each other. We ran a Mann-Whitney U test on the 162 and 179 contingencies where we have data from both pools. Only 4 and 2 contingencies (respectively for generative and preventive conditions) resulted in significant differences (with $p < .05$). None of these differences were significant if Bonferroni correction is applied. Therefore results from these two sources were combined.

Analyses of the dependent chains focused on the final iteration of all chains, as this iteration was most likely to reflect the prior distribution. To test for convergence, we compared the distribution of both ratings as a function of the values used to initiate each chain. ANOVA tests with the initial values of w_0 and w_1 as factors was run for strength estimates of both w_0 and w_1 in each condition. The result showed a statistically significant effect of initial values in the human ratings in the final iteration of both conditions for w_0 but not w_1.

Means and standard deviations from the final iteration of all chains are shown in Table 1. In the generative condition, the w_0 ratings from chains of higher initial w_0 values remains higher than those from chains of lower initial w_0 values ($F(1,286) = 79.568, MSE = 97314, p < 0.001$). On the other hand, there were no significant differences in w_1 ratings between chains of different initial w_1 values ($F(1,286) = 1.000, MSE = 882.0, p = 0.318$). There were no interactions in either chain (in w_0 chain, $F(1,284) = 0.028, MSE = 24.50, p < 0.868$; in w_1 chain, $F(1,284) = 0.8742, MSE = 1073, p = 0.351$). The results were similar in the preventive condition. The w_0 ratings retained influences from initial values ($F(1,334) = 83.474, MSE = 98949, p < 0.001$), but the w_1 ratings are not significantly different ($F(1,334) = 1.831, MSE = 2480.9, p = 0.177$). Again, there were no interactions (w_0 chain: $F(1,332) = 0.557, MSE = 663, p = 0.456$; w_1 chain: $F(1,332) = 1.136, MSE = 1530.0, p = 0.287$).

This pattern of results suggests that the w_1 values had converged while the w_0 values still retained some influence of initialization. Inspection of the data showed that the empirical priors for w_0 has most of its density residing on regions close to 0 or 1 and that individual chains were attracted to these modes, only rarely moving between them, and thus might require longer chain in order to guarantee convergence. However, the final iteration of each chain still gives a reasonably clear picture of the prior. Additionally, since this study focuses on the human judgment of the potential cause C, whose inference is based on w_1, the non-convergence of w_0 does not prevent us from continuing our analysis using this data.

Figure 2 shows the density of the empirical priors based on the responses from the final iteration of all chains, smoothed via kernel density estimation with a bivariate normal kernel (Venables & Ripley, 2002). The generative SS priors gives high probability to regions where only one of w_0 or w_1 is high, while the empirical priors generally prefers w_1 to be high, with the distribution on w_0 being more uniform but has peaks near both 0 and 1. A similar pattern appears in the empirical priors for the preventive case, which is quite different from the prediction under the SS priors in which either w_0 and w_1 are both high or w_0 is high and w_1 is low. Moreover, the prior density is generally lower away from the corners. This characteristic of the empirical prior points to some degree of preference for deterministic causal systems, consistent with previous research on human causal induction (Griffiths & Tenenbaum, 2009; Schulz & Sommerville, 2006).

Comparing models to human judgments

We can now compare Bayesian models based on the three different priors – uniform (Griffiths & Tenenbaum, 2005), sparse and strong (Lu et al., 2008), and the empirical priors estimated by iterated learning – with human judgments of causal strength. Since all three Bayesian models use the same likelihood function and differ only in the priors, the predictions are not radically different. However, the models do differ in

Table 1: Mean and $s.d.$ of the human ratings in the final iteration, separated by initial parameterization.

Chain	Causal direction	Small initial condition	Large initial condition
w_0	Gen.	30.708 (35.498)	67.472 (34.438)
w_0	Gen.	29.577 (34.475)	63.899 (34.384)
w_1	Prev.	81.979 (30.729)	85.479 (28.641)
w_1	Prev.	72.065 (38.289)	77.500 (35.271)

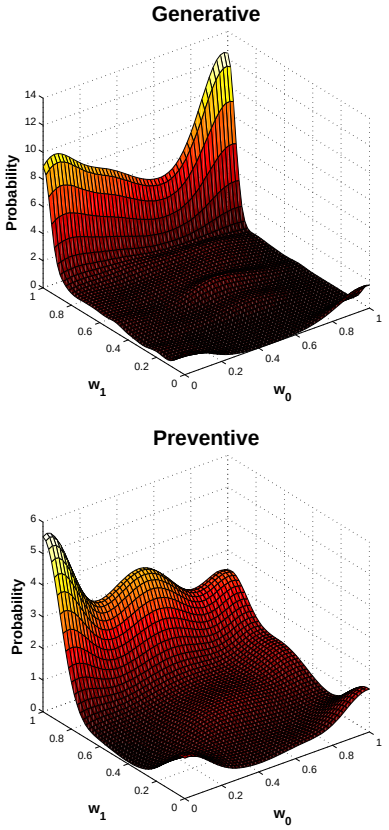

Figure 2: Empirical priors estimated by iterated learning, for both generative and preventive cases.

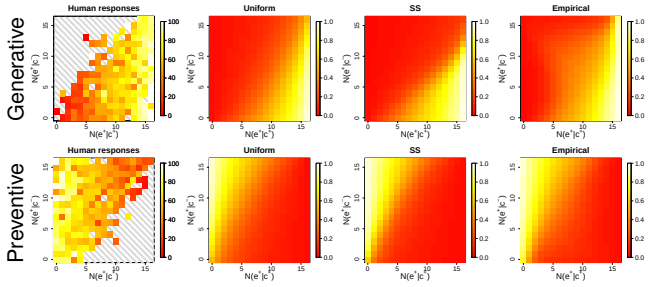

Figure 3: Comparison of human responses with predictions from Bayesian models using uniform, SS, and empirical priors. The grids represent the contingencies with $N(e^+, c^+)$ and $N(e^+, c^-)$ on the horizontal and vertical axes respectively. Cells with gray stripes show contingencies that did not appear in the experiment. Some contingencies are more likely because of sampling.

Table 2: Comparison of model performance on independent chain trials of our experiment.

Causal direction	Metric	Uniform	SS	Empirical
Generative	Correlation (r)	0.7414	0.5620	0.7876
Generative	Adj. RMSD	22.4774	29.6693	15.5900
Preventive	Correlation (r)	0.6544	0.6415	0.6679
Preventive	Adj. RMSD	22.0802	23.3648	19.0560

how they treat cases with middling causal strength, as can be seen by comparing the plots in Figure 3.

The stimuli and, therefore, human data used in most prior studies (e.g. Lu et al., 2008) are not randomly selected and were chosen by researchers in order to compare model performance under specific scenarios. For example, if a researcher is interested in comparing the ΔP model versus the causal power model, often only contingencies that produce radically different predictions under the two models will be selected as stimuli. Although this approach is useful in differentiating models in specific scenarios, it does not necessarily reflect causal induction more generally. We opted for a more general approach and used the responses produced in the independent chains from our experiment, where trials were generated using uniformly distributed w_0 and w_1 values and therefore covered a wide range of contingencies.

The performance of the models was evaluated using Pearson's correlation coefficient (r) and an adjusted root-mean-square deviations (RMSD) score. The correlation r compares the mean human judgment with model predictions. All three models performed quite well based on the Pearson's correlation coefficient (r) with the empirical model having the highest correlation. The results are shown in Table 2. Overall, the r metrics in this experiment are lower compared to those in prior studies (e.g. Lu et al., 2008) where values of r over 0.95 are not uncommon. This is expected because of our more comprehensive data set, with many different contingencies, and as a result, higher variability in mean human judgments. For example, there were eight contingencies (in each causal direction) in Experiment 1 of Lu et al. (2008) whereas there were 184 and 191 contingencies in our experiment, in the generative and preventive case respectively.

In order to take into account the differences in the sample size of the human responses, we also evaluated the models using adjusted RMSD. We calculated this score as $\textit{Adjusted RMSD} = \sqrt{\sum_i \frac{(\hat{x}_i - x_i)^2}{se_i} / \sum_j \frac{1}{se_j}}$ where x_i are the means of the human responses, $\hat{x}_i$ are model predictions, and se_i are the standard errors. This metric assigns higher weight to contingencies where more human data are available (thus lower sampling error) and where variability is smaller. Contingencies where the standard error could not be computed were omitted from the analysis. With this metric, the empirical priors again made better predictions than the other two models, particularly in the generative direction. Figure 4 compares the performance of models using the SS and empirical priors in terms of adjusted RMSD at every set of contingencies that appeared in the experiment.

Discussion

We have presented the first direct estimate of human priors on the strength of causal relationships, using iterated learn-

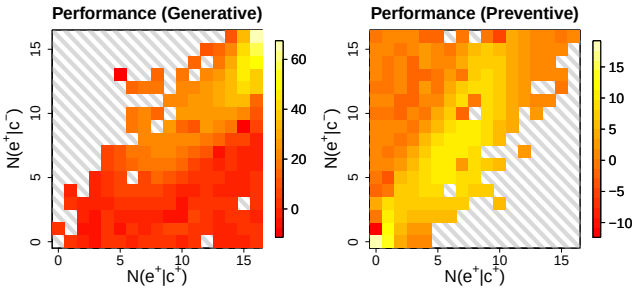

Figure 4: Comparison of model predictions. These two plots compare the performance of the Bayesian model with the empirical priors against that with the SS priors. The grids represent the contingencies with $N(e^+, c^+)$ in x-axis and $N(e^+, c^-)$ in y-axis. The value at each grid is the difference of the error of the models (compared to the mean human result). Positive values (lighter grids) represent better performance for the empirical prior model; negative values represent better performance for the SS prior model. Only data from the independent chains are plotted here.

ing. The resulting empirical priors were markedly different from previously proposed theory-based priors. We also found that the empirical priors predict human judgments better than these previously proposed priors. However, there are a number of important issues that need to be explored in future work, including giving a more comprehensive characterization of priors across different causal scenarios, dealing with more complex causal systems, and determining an objective method for evaluating models of causal induction.

The gene expression cover story we used in our experiment is similar to cover stories used in numerous prior experiments on causal induction (Lu et al., 2008; Griffiths & Tenenbaum, 2005; Lober & Shanks, 2000). These medical cover stories are used because they provide plausible causal relationships between variables, and the functional form of these relationships is simple. However, it is possible that prior knowledge might influence the form of the priors that people use when reasoning about this particular scenario, such that the empirical priors we estimated might not generalize well to other causal domains. It is also possible that cultural differences might influence the form of the priors. Having established that iterated learning can be used to investigate priors on causal strength, we anticipate that this approach can be used to explore how priors vary across scenarios and cultures.

By focusing on the simplest possible causal structure, our study did not address how people reason about causal systems where more than one non-background cause is present. This reflects a general phenomenon in the study of human causal induction, where investigation of systems involving multiple causes is rare. However, some recent studies have examined how people estimate the strength of multiple causes (e.g., Novick & Cheng, 2004). Adapting the methods used in these studies to an iterated learning setting might provide a way to investigate whether the priors that people adopt depend on the complexity of the causal system.

Finally, our results also raise questions concerning how models of causal induction should be compared. Most past causal inference literature evaluate model fit by comparing model predictions with human responses at specific contingencies. These contingencies were usually chosen to highlight the different predictions made by the models of interest. The independent chains used in our experiment provided a picture of human causal strength judgments for a remarkably wide range of contingencies. Establishing a broad database for objectively comparing models of human causal induction is an important challenge for future research.

Acknowledgments. This work was supported by grant number FA-9550-10-1-0232 from the Air Force Office of Scientific Research, and the McDonnell Causal Collaborative.

References

Buehner, M., & Cheng, P. (1997). Causal induction: The power PC theory versus the Rescorla-Wagner model. In *Proceedings of the nineteenth annual conference of the cognitive science society: August 7-10, 1997, stanford university* (p. 55).

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Griffiths, T., Christian, B., & Kalish, M. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*(1), 68–107.

Griffiths, T., & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science: A Multidisciplinary Journal*, *31*(3), 441–480.

Griffiths, T., & Tenenbaum, J. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T., & Tenenbaum, J. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716.

Hume, D. (1739/2004). *An Enquiry Concerning Human Understanding*. Dover Publications.

Kalish, M., Griffiths, T., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, *5*, 102-110.

Lober, K., & Shanks, D. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review*, *107*(1), 195–212.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P., & Holyoak, K. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955–984.

Novick, L., & Cheng, P. (2004). Assessing Interactive Causal Influence. *Psychological Review*, *111*(2), 455–485.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and children's inferences about unobserved causes. *Child Development*, *77*, 427-442.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search*. The MIT Press.

Venables, W., & Ripley, B. (2002). *Modern applied statistics with S*. Springer-Verlag.

Ward, W., & Jenkins, H. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology*, *19*(3), 231–241.