Cognitive Psychology 76 (2015) 1–29

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Identifying expectations about the strength of causal relationships

Saiwing Yeung [a,*], Thomas L. Griffiths [b]

[a] Institute of Education, Beijing Institute of Technology, China
[b] Department of Psychology, University of California, Berkeley, United States

ARTICLE INFO

Article history:
Accepted 3 November 2014

Keywords:
Causal reasoning
Bayesian models
Computational modeling
Iterated learning

ABSTRACT

When we try to identify causal relationships, how strong do we expect that relationship to be? Bayesian models of causal induction rely on assumptions regarding people's a priori beliefs about causal systems, with recent research focusing on people's expectations about the strength of causes. These expectations are expressed in terms of prior probability distributions. While proposals about the form of such prior distributions have been made previously, many different distributions are possible, making it difficult to test such proposals exhaustively. In Experiment 1 we used iterated learning—a method in which participants make inferences about data generated based on their own responses in previous trials—to estimate participants' prior beliefs about the strengths of causes. This method produced estimated prior distributions that were quite different from those previously proposed in the literature. Experiment 2 collected a large set of human judgments on the strength of causal relationships to be used as a benchmark for evaluating different models, using stimuli that cover a wider and more systematic set of contingencies than previous research. Using these judgments, we evaluated the predictions of various Bayesian models. The Bayesian model with priors estimated via iterated learning compared favorably against the others. Experiment 3 estimated participants' prior beliefs concerning different causal systems, revealing key similarities in their expectations across diverse scenarios.

[*] Corresponding author at: Institute of Education, 5 South Zhongguancun Street, Haidian District, Beijing 100081, China.
E-mail address: saiwing.yeung@gmail.com (S. Yeung).

http://dx.doi.org/10.1016/j.cogpsych.2014.11.001
0010-0285/© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Inferring the relationship between causes and effects is an important skill that people rely on every day in order to understand the structure of their environment. Psychological models of human causal induction have often focused on the role of associative learning—how people's judgments might be related to the number of instances of an effect occurring in the presence and absence of a cause (e.g., Cheng, 1997; Shanks, 1995; Ward & Jenkins, 1965). Recent work has explored how ideas from Bayesian statistics might help to explain people's intuitions, 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 influence of prior knowledge (Griffiths & Tenenbaum, 2009).

A key part of Bayesian models of causal induction is the assumptions they make about the expectations people have about causal relationships. These assumptions are expressed in the form of *prior distributions* (often shortened to just *priors*), and reflect the expectations of the learners about causal systems prior to seeing any data. Bayesian models of human causal induction focus especially on the learners' prior beliefs about the *strength* of causal relationships, and two different specifications have been proposed thus far—using either a non-informative prior that makes no assumptions about the prior beliefs (Griffiths & Tenenbaum, 2005), or a prior based on theoretical assumptions about people's inductive biases (Lu et al., 2008). In this paper we present a new approach to identifying the expectations people have about causal systems, using the technique of *iterated learning* (Griffiths & Kalish, 2007; Griffiths, Christian, & Kalish, 2008). This technique allows us to estimate the values of unobservable cognitive constructs, such as people's priors, and produced predictions about human judgments that are more accurate than previous methods.

The plan for the rest of the paper is as follows. In the next section we summarize relevant previous work on Bayesian models of human causal induction. We then introduce the basic ideas behind iterated learning, and apply it in Experiment 1 to empirically estimate participants' prior beliefs about causal systems. Next we discussed a potential issue in previous work on human causal induction—models of causal induction are often evaluated based on small sets of stimuli that have certain specific characteristics. To address this issue, in Experiment 2 we collected a large benchmark data set against which the performance of different models can be better compared. We followed up with an investigation on prior beliefs about several different causal systems in Experiment 3. Finally we conclude with a discussion of the implications of these results for understanding causal induction as well as some possible future directions.

2. Bayesian models of human causal induction

Historically, there have been two major approaches to psychological theories of causal induction (Newsome, 2003). The mechanism-based approach focuses on understanding how knowledge about the causal mechanisms influences reasoning. Here learners attempt to discover a process in which causal power possessed by entities is transmitted to generate the event (Ahn, Kalish, Medin, & Gelman, 1995). In contrast, the covariation-based approach focuses on how people use the contingency about cause and effect to identify causal relationships (Cheng, 1997; Shanks, 1995). In particular, much psychological research on causal induction has focused on the problem of *elemental causal induction*: given a number of observations of two binary variables with a plausible causal relationship, how do people assess the relationship between them? For example, we can imagine a scientist studying the effect of a certain chemical on clovers. While most clovers have three leaflets, some have four, and the scientist wants to know whether the presence of the chemical has the power to increase the proportion of four-leaf clovers. By planting a number of clover plants and applies this chemical on some of them, he can use the resulted contingency data—number of clover plants exposed to or not exposed to the chemical, number of four-leaf clovers resulted in each case—to evaluate this potential causal relationship.

More recently, researchers have proposed various models based on Bayesian statistics. While these models rely on covariation data, they explicitly represent the learners' subjective beliefs about the causal systems, and formally express how they update their beliefs after observing data. These models have been found to be able to make good predictions about human judgments, and could explain effects not predicted by previous models (Griffiths & Tenenbaum, 2005; Lu et al., 2008). This paper will focus on these Bayesian models.[1]

2.1. Bayesian inference

Bayesian models of causal induction are special cases of the more general approach of modeling human inductive inferences as applications of Bayesian statistics (Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). The principles of Bayesian inference indicate how a rational learner should update his or her beliefs in light of evidence. In the context of causal induction, this approach involves formally specifying a learner's a priori beliefs concerning a causal system in terms of prior probability, and a likelihood function that describes the likelihood of observing certain data given these prior beliefs. The prior probability and the likelihood function combine to give the posterior probability, which represents the learner's updated beliefs concerning the causal system.

Formally, we represent a learner's a priori subjective beliefs concerning a causal system using a set of hypotheses $\mathcal{H}$, in which each hypothesis $h \in \mathcal{H}$ has a prior probability $P(h)$ that reflects the agent's degree of belief in the hypothesis being true.[2] Bayes' rule indicates that the degree of belief in each hypothesis after observing data d is given by

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}$$

where $P(d|h)$ is the likelihood and $P(h|d)$ is the posterior probability. The posterior probability thus represents the learner's belief after the observations, and can be compared to empirical data as a way to evaluate to what extent the proposed model can approximate people's computations in causal induction.

Because these Bayesian models rely on computations concerning people's beliefs, these beliefs need to be formalized mathematically. However, formally specifying a prior distribution that captures the expectations of humans (or even just a select group of a more homogeneous population) is particularly difficult, and will be the main concern of the present paper.

2.2. Modeling causal induction

Griffiths and Tenenbaum (2005) presented a Bayesian analysis of causal induction, in the spirit of Marr (1982) and Anderson (1991). This analysis used causal graphical models to formulate the problem of causal induction and outlined a rational solution to this problem based on Bayesian inference. Causal graphical models are probabilistic models in which graphs are used to denote the causal relationships between variables (Pearl, 2009; Spirtes, Glymour, & Scheines, 2001). In these graphs, nodes represent variables and edges represent the causal connections between those variables. An elemental causal system is formalized using three variables: the background cause B, the candidate cause C, and the effect or outcome E (Fig. 1). In the scientist example discussed earlier, we have an observed effect E (four-leaf clovers), which might be caused by the candidate cause C (the chemical), or the background cause B (representing any other causes that are not of interest). We assume that both B and C can cause E, and this relationship is expressed by edges going from both B and C to E. We further 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. In the generative case, either B or C can cause E; in the

[1] A different class of covariation-based computational models of causal reasoning focuses on deriving algebraically from a contingency table people's evaluation of causal relationships. These models include ΔP (Ward & Jenkins, 1965), causal power (Cheng, 1997), EI rule (Perales & Shanks, 2007), etc. A comprehensive treatment of these models appears in Perales and Shanks (2007).

[2] In this paper we use the convention of using $p()$ to represent continuous distributions, and $P()$ to represent discrete distributions and proportions.

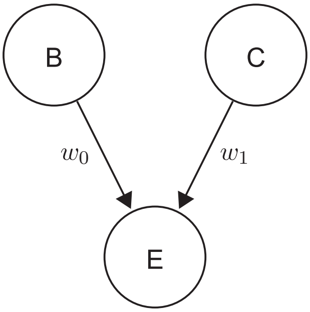


Fig. 1. Causal graphical model. The elemental model involves three variables: *B* the background cause, *C* the candidate cause, and *E* the effect of interest. While *B* is always present, *C* and *E* might be present or absent. *B* and *C* are assumed to be independent. *B* is always generative; *C* could be generative or preventive. Independently, *B* and *C* have probabilities w_0 and w_1, respectively, of influencing *E*. Their joint functional form is formalized using noisy-OR in the generative case and noisy-AND-NOT in the preventive case.

preventive case, only *B* can cause *E*, while *C* may suppress *E*. Finally, *E* cannot occur unless *B* or *C* caused it. 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 strengths of *B* and *C* as w_0 and w_1 respectively.

Although the graph structure specifies the causal relationships among variables, the exact nature of those relationships is not clear without specifying their *functional forms*. Functional forms are the formal specifications associating the states of causes to the probability of different outcomes, and are a key component in the likelihood function. To define the functional form of a causal relationship, we need to make the distinction between *generative* and *preventive* causes—generative causes have the power to produce the effect, whereas preventive causes have the power to prevent the effect from occurring. While many early theories of causal induction did not make an explicit distinction between these kinds of causal relationships, more recent research suggested that people reason about generative and preventive causal relationships differently (Dennis & Ahn, 2001; Williams & Docking, 1995). As a consequence, recent models of causal induction have distinct variants for characterizing judgments about generative and preventive causes (Cheng, 1997; Griffiths & Tenenbaum, 2005).

In previous models of causal induction, noisy-OR (for generative causes) and noisy-AND-NOT (for preventive causes) parameterizations have been used to characterize the functional form of causal relationships (Cheng, 1997; Griffiths & Tenenbaum, 2005). Noisy-OR gives the probability of observing the effect *E* as

$$p(e^+|c; w_0, w_1) = 1 - (1 - w_0)(1 - w_1)^c \tag{1}$$

where *c* is a binary value representing the presence or absence of *C* with $c = 1$ for c^+ (presence of the candidate cause) and $c = 0$ for c^- (absence of the candidate cause), whereas noisy-AND-NOT gives

$$p(e^+|c; w_0, w_1) = w_0(1 - w_1)^c \tag{2}$$

Again, w_0 and w_1 represent the strength of *B* and *C* respectively. These formulations reflect the above assumptions we made concerning elemental causal induction.

Recent work by Lu et al. (2008) demonstrated how this framework could be used to make accurate predictions about people's estimates of the strength of causal relationships. Here the hypotheses space is represented by the prior probability over w_0 and w_1. For any particular value of w_0 and w_1, the likelihood function for the observed contingency data *d* is

$$p(d|w_0, w_1) = \prod_{e,c} p(e|c; w_0, w_1)^{N(e,c)} \tag{3}$$

where $p(e|c; w_0, w_1)$ is given by the noisy-OR or noisy-AND-NOT as specified above. $N(e, c)$ represents the frequencies of events associated with different combinations of presence and absence of the effect and cause. Each of e and c can be either $^+$ or $^-$, representing presence or absence, respectively.[3]

We can thus compute a posterior distribution over w_0 and w_1 given d by applying Bayes' rule, with

$$p(w_0, w_1|d) \propto p(d|w_0, w_1)p(w_0, w_1) \tag{4}$$

where $p(w_0, w_1)$ represents the prior probability of a (w_0, w_1) pair. While the posterior probability represents a learner's updated beliefs over the entire hypothesis space, we can find point estimates of the posterior of w_0 and w_1 by computing the mean of the posterior distribution. For example, posterior mean of w_1 computed based on

$$\bar{w}_i = \int_0^1 w_i p(w_0, w_1|d)\, \mathrm{d}w_i \quad \text{for } i \in 0, 1 \tag{5}$$

has been found to be able to approximate closely people's judgments about the strength of causes (Lu et al., 2008).

2.3. Priors on causal strength

Intuitively, a prior distribution on causal strengths encapsulates the learner's expectations about the strengths of causes, which corresponds to the probability with which they bring about (or prevent) the effect. In an elemental causal induction model, there are two relevant causes—the background cause B and the candidate cause C—and therefore their joint prior probability is the key to understanding human causal induction. Griffiths and Tenenbaum (2005) assumed uniform priors—the simplest non-informative prior—on both variables in their structure-learning model. Under this parameterization, equal value is assigned to all possible values of w_0 and w_1, reflecting indifference about their values (Jaynes & Bretthorst, 2003). In spite of its simplicity and lack of free parameters, the uniform model provided a good fit to human judgments about causal structure.[4]

Lu et al. (2008) applied this approach to model how people reason about strengths of causes and argued that human causal induction is better explained using a model that incorporates *generic priors*—a theoretically driven set of prior distributions that are derived from assumptions about the abstract properties of a system. They argued that people favor necessary and sufficient causes, and therefore people consider causal models that have fewer causes (Chater & Vitányi, 2003) and have no complex interactions (Novick & Cheng, 2004) to be more likely. As elemental causal induction involves only one background cause and one candidate cause, this suggests that people prefer a causal system in which only one of the two causes is necessary and sufficient to generate the effect. Based on these arguments, Lu et al. (2008) specified the *sparse and strong* prior (SS prior) as

$$p(w_0, w_1) \propto e^{-\alpha(w_0+1-w_1)} + e^{-\alpha(1-w_0+w_1)} \tag{6}$$

in the generative case and

$$p(w_0, w_1) \propto e^{-\alpha(1-w_0+1-w_1)} + e^{-\alpha(1-w_0+w_1)} \tag{7}$$

in the preventive case. α in the formulae is a free parameter that represents how strongly one believes that causes are sparse and strong. If α is set to 0 then the SS priors are identical to a uniform distribu-

[3] Following Griffiths and Tenenbaum (2005), we use a notation that semantically encodes the presence or absence of the effect and cause—e and c represent the effect (or outcome) and candidate cause, and the superscripts $^+$ and $^-$ represent their presence and absence, respectively. For example, the number of instances in which the effect is present but the cause is absent would be represented by $N(e^+, c^-)$. In some cases, it is convenient to use the relative frequencies of the outcomes. These can be obtained from the $N(e, c)$ values. For example, the relative frequency of the effect given the presence of the cause is $P(e^+|c^+)$ and can be calculated by $\frac{N(e^+,c^+)}{N(c^+)}$, where $N(c^+)$ denotes the number of cases in which the cause is present, regardless of whether the effect is present or not. Similarly, $N(c^-)$ denotes the number of cases in which the cause is absent and the relative frequency $P(e^+|c^-)$ and can be calculated by $\frac{N(e^+,c^-)}{N(c^-)}$.

[4] For the ease of exposition, in the remainder of this paper, we will refer to Bayesian models that use different priors as different models, such as "the uniform model" instead of "the Bayesian model using a uniform prior".

tion. Lu et al. (2008) fixed α at 5 in their analysis, based on an informal grid search using a prior data set.

The SS priors are plotted as two-dimensional histograms in Fig. 2. Analogous to one-dimensional histograms, these figures indicates the probability density at different values of the underlying variables (w_0 and w_1). From these figures we can observe the main features of the SS priors. In the generative case, the probability is high when one of the causes, either B or C, is very strong and the other is very weak; in the preventive case, the probability is high when B is very strong and C is either very strong or very weak.

The SS model provided a good fit to human causal strength judgments, and produced predictions consistent with qualitative effects that a model with a uniform prior does not produce (Lu et al., 2008). This finding suggests that exploring the space of possible prior distributions might be a valuable way of gaining deeper insight into human causal induction, as well as better predictions of human behavior. However, prior distributions on causal strength could take any form that corresponds to a probability distribution on two variables ranging from 0 to 1, and need not correspond to a distribution from a simple parametric family. This means that the SS priors represent only one possibility out of uncountably many. Evaluating all possible priors is thus infeasible. Ideally, the prior distribution should be estimated without having to make a priori assumptions about its form. In the next section we explore how the technique of iterated learning might be used to resolve this issue.

3. Iterated learning as a method for estimating priors

Iterated learning was originally proposed as a model of cultural transmission of languages (Kirby, 2001). It refers to a process in which a sequence of agents each learns from data generated by the previous agent. In the simplest such model 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.

Formally, the n-th agent in the chain observes data $d^{(n)}$ and forms a hypothesis $h^{(n)}$ about the process that generated those data, then goes on to generate data $d^{(n+1)}$, which is given to the next agent, and so on. This iterative procedure defines a Markov chain—a sequence of random variables in which

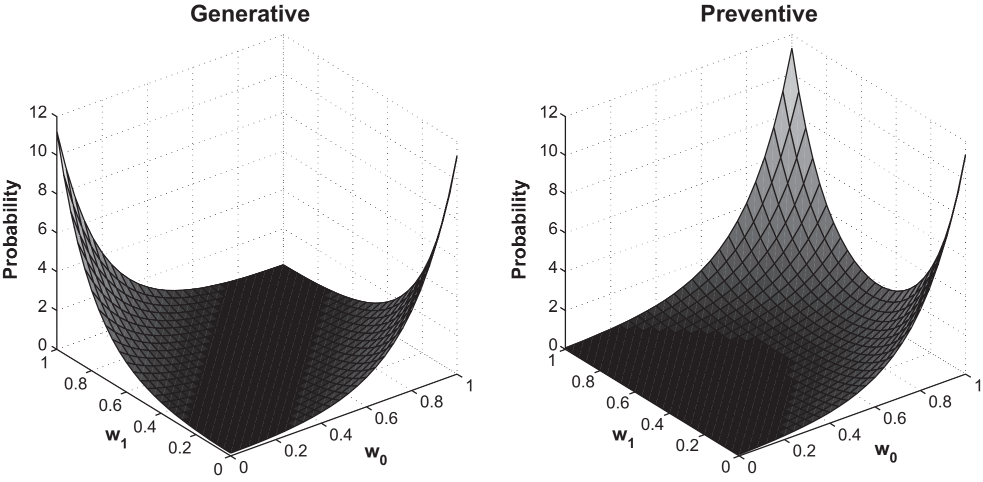


Fig. 2. Sparse and strong (SS) priors (Eqs. (6) and (7)) proposed by Lu et al. (2008). Each panel shows the prior distribution for either the generative or the preventive cause. The two horizontal axes correspond to the strength of the background cause (w_0) and the candidate cause (w_1). The vertical axis indicates the relative probability density at each (w_0, w_1) pair. Note that the density was normalized before plotting (i.e., the volume under the surface is 1). Therefore the density values do not match those given directly by the formulae.

each variable's transition depends only on the current state—on hypothesis-data pairs. That is, each step of the Markov chain here consists of the hypothesis about the causal system and the data produced, and the hypothesis-data pair at each step depends only on its immediate previous iteration.

Furthermore, 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 forms a *Gibbs sampler* for the joint distribution $p(d,h) = p(d|h)p(h)$ (Griffiths & Kalish, 2007). A Gibbs sampler is a specific kind of Markov chain, commonly used by computer scientists and statisticians to generate samples from a multivariate probability distribution that normally might be difficult to sample from Gilks, Richardson, and Spiegelhalter (1996).

More formally, a Gibbs sampler provides a way to sample from the joint probability distribution for a set of variables $X = (x_1, \ldots, x_n)$. This procedure involves an initial stage and a number of iterative steps. We first start with some initial values for each of the variables, and then at each step, sample the value of one of the variables in the target distribution by drawing a sample from the distribution of this variable, conditioned on the current values of all other variables. That is, at each iteration, the algorithm cycles through the variables and draws a new value for each variable x_i from the conditional distribution $p(x_i|X_{-i})$. This process constructs a Markov chain that, under mild assumptions, will ultimately converge to its stationary distribution $p(X)$, at which point samples can be taken as an approximation of the target distribution.

In an iterated learning experiment, the procedure of alternately sampling hypotheses and data mimics the structure of the Gibbs sampler, allowing the same convergence results to be applied (Griffiths & Kalish, 2007). In this case, samples of h after convergence can be taken as samples generated from the prior distribution $p(h)$. In other words, the probability that a participant selects a particular hypothesis will converge to her prior. Carrying out this process of iterated learning in experiments thus enables us to empirically estimate people's priors. This result holds true regardless of what data are provided to the subjects in the first iteration.

The convergence of iterated learning to the prior suggests that it can be used as a method for exploring people's expectations about different hypotheses. Previous experiments using iterated learning, with data being passed between people, support this idea: functions (Kalish, Griffiths, & Lewandowsky, 2007), concepts (Griffiths et al., 2008), and color terms (Xu, Dowman, & Griffiths, 2013) transmitted through iterated learning quickly converge to forms that are consistent with priors established in previous research. Moreover, there is no need for data to be transmitted between people for this to occur—a feedback process that has the appropriate statistical structure can be established for single individuals. In these within-subjects iterated learning experiments, participants form hypotheses about data that are generated based on their own responses on previous trials. This experimental design has previously been used to explore people's inductive biases in concept learning and memory, producing equivalent results to a between-subjects design (Griffiths et al., 2008; Xu & Griffiths, 2010).

In this paper we use the iterated learning approach in experiments to directly estimate the prior distributions of the strength parameters (w_0 and w_1) in people's subjective beliefs concerning causal systems. This method could inform recent discussion about appropriate prior distributions to use in Bayesian models of causal learning, and result in more accurate predictions of human judgments. In particular, iterated learning allows us to identify people's expectations about the strength of causal relationships without needing to make any assumptions about the form of the corresponding prior distributions. We will first demonstrate how this approach can be used to identify participants' prior beliefs in a simple biological scenario.

4. Experiment 1: Using iterated learning to estimate human priors on causal strength

The objective of Experiment 1 is to estimate people's priors in elemental causal induction using the technique of iterated learning. The priors obtained, representing the experiment participants' expectations about the strengths of the causes, can be used to evaluate previous claims about human prior beliefs.

4.1. Methods

4.1.1. Participants

A total of 432 participants were recruited from the University of California, Berkeley subject pool and Amazon Mechanical Turk (MTurk). Collecting data from MTurk in addition to a more traditional university subject pool allowed us to obtain data that are more representative of the entire population than just the university sample, and to run a larger scale experiment. We recruited only MTurk workers from within the United States. Recent studies have shown MTurk to be a reliable source of experimental data (Buhrmester, Kwang, & Gosling, 2011; Sprouse, 2011) and many classical experiments were successfully replicated using this service (Crump, McDonnell, & Gureckis, 2013). Participants from the university subject pool received course credit, while MTurk participants received a payment of US$1.[5] Only data from participants who completed at least 95% of the trials were included in the analysis. In each of the generative or preventive condition, there were 36 participants from the university subject pool and 180 from MTurk.

In the university subject pool, 55% of participants were female and the mean age was 20.1. In the MTurk subject pool, 49% were female. With MTurk participants, we requested age information in brackets. The biggest age group was 23–35, with 50% of our participants, and 78% had at least some college education. The demographics of the MTurk subject pool were thus similar to our university subject pool, other than their slightly higher age, and are also commensurate with other published reports on the demographics of MTurk workers (Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010).

4.1.2. Stimuli and procedure

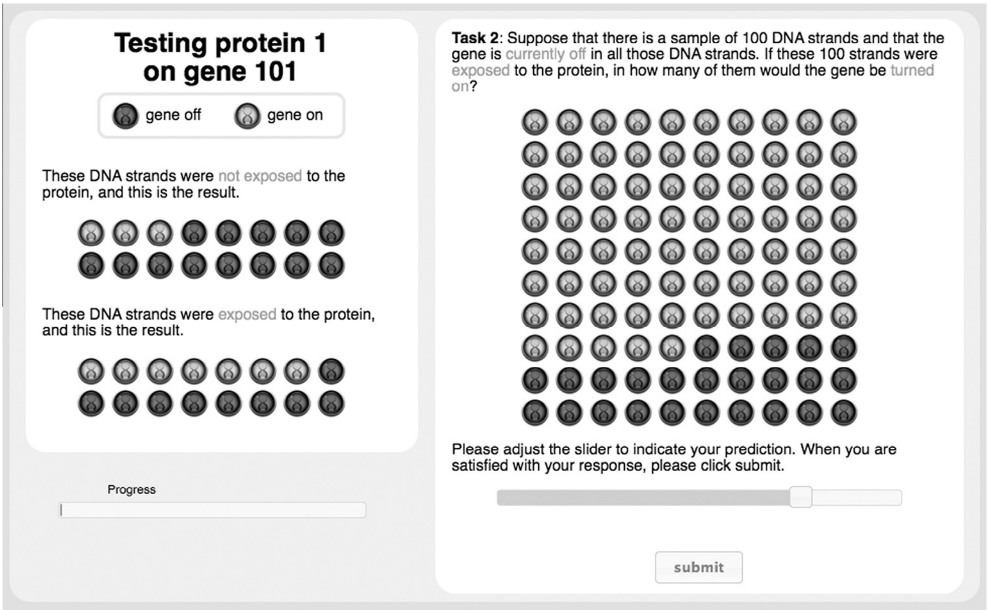
The experiment was run in a web browser. We used a biological cover story that most closely resembled one that was used in Lu et al. (2008), but cover stories of similar themes have been used in many previous studies of causal induction (e.g., Buehner, Cheng, & Clifford, 2003; Griffiths & Tenenbaum, 2005; Novick & Cheng, 2004). We chose to use a similar cover story for two reasons. First, as we are using the novel experimental technique of iterated learning, having a well-tested cover story and associated stimuli could spare us from having to evaluate multiple new experimental elements at the same time. Second, reusing the same materials allows better comparison between different models, as most of these models were originally evaluated based on experiments using similar stimuli.

The experiment was presented in the context of a bio-technology company testing the influence of various proteins on gene expression. In the *generative* condition, participants read:

> 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 a process that 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 predictions concerning the effect of these proteins on gene expression.
>
> There are a number of trials in this experiment and each trial involves a different gene and a different protein. In each trial, you will be given information about some past results involving this gene/protein pair and you will be asked to make some predictions based on these information. The past results consist of two samples: 1) a sample of DNA fragments that had not been exposed to the protein, and 2) a sample of DNA fragments that had been exposed to the protein. The number of DNA fragments that resulted in gene expression in each of these samples will be shown to you. Because there are many causes of gene expression, other factors besides the presence or absence of the protein might play a role in whether the gene is expressed or not.

Participants then received instructions familiarizing them with the controls that they would use in the experiment. Only one practice trial was used in order to minimize effect of practice on participants' expectations, as each practice trial involves a certain set of contingency that could impart influence on participants' beliefs about the novel causal system. Each trial was presented on a separate

[5] The amounts of payment for experiments in this paper were comparable to similar experiments on MTurk (Mason & Suri, 2012).



Testing protein 1 on gene 101

gene off gene on

These DNA strands were not exposed to the protein, and this is the result.

These DNA strands were exposed to the protein, and this is the result.

Progress

Task 2: 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?

Please adjust the slider to indicate your prediction. When you are satisfied with your response, please click submit.

Fig. 3. Screenshot of the *generative* condition of the experiment. The participant is assessing the strength of the cause C, w_1. In this particular trial, $\frac{N(e^+,c^-)}{N(c^-)} = \frac{3}{16}$ and $\frac{N(e^+,c^+)}{N(c^+)} = \frac{7}{16}$.

screen. Fig. 3 shows a screenshot of the experiment. In each trial participants saw data in the form of two samples, one that was not exposed to the protein ($N(e, c^-)$) and one that was ($N(e, c^+)$). The data were presented graphically, displaying the total number of DNA fragments in each sample as well as the number that expressed the gene, thus providing complete contingency data.

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

> Suppose that there is a sample of 100 DNA fragments and these fragments 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 fragments and that the gene is currently off in all those DNA fragments. If these 100 fragments were exposed to the protein, in how many of them would the gene be turned on?

These questions were similar to those used in previous research on causal strength judgments for eliciting judgments of w_0 and w_1, respectively (e.g., Lu et al., 2008). Participants responded using a slider, the initial value of which was drawn from a uniform distribution between 0 and 100. Live feedback showing the proportion of expressed genes was shown as the slider was moved. Past studies have shown that graphically presented information can better convey probabilistic data (Edwards, Elwyn, & Mulley, 2002). Participants could adjust the slider until they were satisfied with their response, before clicking the submit button to record their response and continue to the next trial. The instructions for the *preventive* condition were similar, except that for the second judgment, the instructions were:

> Suppose that there is a sample of 100 DNA fragments and that the gene is currently on in all those DNA fragments. If these 100 fragments were exposed to the protein, in how many of them would the gene be turned off?

A within-subjects iterated learning design was used. There were four chains each with 12 iterations, for a total of 48 trials. The data for the first iteration of each chain were generated based on the initial values of $(w_0, w_1) = (0.3, 0.3), (0.3, 0.7), (0.7, 0.3)$, and $(0.7, 0.7)$. These values were chosen to be distinct so that the responses from different chains could be used to diagnose convergence.

In each trial a set of contingency data consisting of 16 cases in which the cause was present and 16 in which the cause was absent was shown. The number of times that the effect occurred was generated via a binomial draw with parameter $p(e^+|c; w_0, w_1)$, using the noisy-OR functional form (Eq. (1)) in the generative case and noisy-AND-NOT (Eq. (2)) in the preventive case. In all trials other than the first one in each chain, 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 same chain would be drawn using $w_0 = f_0/100$ and $w_1 = f_1/100$. Additionally, the order of the trials was randomized after stratification by iteration. For example, the first four trials for each participant correspond to the first iteration of each of the four chains, but the order among these four trials was randomized for each subject. Similarly, the fifth to the eighth trials correspond to the second iteration for each of the four chains, and so on.

4.2. Results

We first compared the results between the university and MTurk subject pools. The numbers of contingencies in the data were not fixed because of the stochasticity in the sampling of stimuli. There were 181 and 165 contingencies in the generative and preventive conditions respectively, for which there were data in both pools. We ran a Mann–Whitney U test on these contingencies and found that there were significant differences (with $p < .05$) between the two subject pools in 10 and 5 contingencies respectively. None of these differences were significant if Bonferroni correction is applied. Therefore results from these two sources were combined.

4.2.1. Analysis of convergence

Before we compute the prior distributions using the empirical data, we tested whether the data indeed have reached convergence. Because of the divergent initial values of the four chains, we expected the human judgments for both w_0 and w_1 to be significantly different across chains of different initial values in the first few iterations. As the iterated learning experiment progresses, the hypotheses-data pair should eventually converge to the prior distribution and will no longer be significantly different across chains. We will compare the human judgments between initial value conditions to diagnose whether the chains have converged. This convergence diagnosis method is similar to those used in Markov chain Monte Carlo methods in statistics (Gilks et al., 1996).

In the first iteration the judgments for both w_0 and w_1 were significantly different, indicating that, as we had expected, the chains have not converged. The result was quite different for the final (12th) iteration. In the generative condition, there were no significant differences between chains with different initial w_1 values ($F(1, 860) = 0.47, MSE = 519.6, p = 0.49$). However, the w_0 judgments from chains of higher initial w_0 values remained higher than those from chains of lower initial w_0 values ($F(1, 860) = 58.31, MSE = 79,945, p < 0.001$). The results were similar in the preventive condition. The w_1 judgments from chains with different initial values were not significantly different ($F(1, 860) = 0.06, MSE = 65.0, p = 0.81$), while the w_0 judgments were ($F(1, 859) = 61.00, MSE = 83,128, p < 0.001$). The means and standard deviations of the chains at the final iteration are shown in Table 1.

The significant effect of initial values for w_0 but not w_1 in both conditions indicates that the w_1 chains had converged while the w_0 chains still retained some influence from the initial values. However, we believe that samples from the final (12th) iteration give a fairly good approximation of the prior distributions for two reasons. First, while the w_0 chains did not converge, they did move in the expected directions towards converging. It is likely that they would converge given an experiment with more iterations.[6] Second, as the focus of this experiment is on human judgments of the strength of

[6] While it is possible to collect human data for more iterations until w_0 converges, a mathematical analysis suggests that this might not be practical. We conducted a simulation using the uniform prior to estimate how long it would take to converge, using the same initial w_0 and w_1 values as the experiment. Using the same criteria for convergence as in the experiment, we found that on average, it took about 30 iterations for w_0 to converge. With four chains (to account for low and high initial values of w_0 and w_1), it would take a total of 120 trials, which might be too many that boredom and fatigue could become a factor.

the candidate cause, which is most strongly influenced by the prior distribution of w_1, the non-complete convergence of w_0 presents a more minor issue than if w_1 did not converge. In fact, the w_1 chains in both causal directions converged by the third iteration, suggesting that humans have fairly strong priors about the strength of the candidate cause.

4.2.2. Estimation of the empirical priors and discussion

To construct the empirical priors, we treated judgments of w_0 and w_1 in the final iteration as samples from participants' prior distributions. Because there were four chains per participant, each participant contributed four data points, each in the form of a (w_0, w_1) vector. We aggregated the (w_0, w_1) vectors from all participants and performed a two-dimensional density estimation. The top row of Fig. 4 shows a smoothed estimate of the density of the empirical priors over w_0 and w_1. The smoothing was performed via kernel density estimation with a bivariate normal kernel (Venables & Ripley, 2002).

The resulting empirical prior distributions, shown in Fig. 4, have complex forms that do not correspond to parametric probability distributions. Moreover, their forms are quite different from previous proposals about people's prior distributions, such as the SS priors. Some major differences between the empirical and the SS priors can be observed based on visual inspection. First, the SS theory suggests that people expect either the candidate cause or the background cause to be strong. In contrary to this assumption, in the empirical priors, the density is much higher at regions where w_1 is high, in both causal directions. In other words, participants have a strong prior for a strong candidate cause, regardless of whether the background cause is strong or not.

Second, the SS priors are relatively straightforward probability distributions. Here probability is the highest at the two peaks (which have the same density), and from there the probability goes down strictly. Moreover, in both the generative and preventive cases, the two non-peak corners have the lowest prior probability of the entire distribution. For example, in the generative case, the peaks are at $(w_0 = 0, w_1 = 1)$ and $(1, 0)$ while the other two corners of $(0, 0)$ and $(1, 1)$ have the lowest prior probability of the entire space, and similarly in the preventive case. In contrast, the empirical priors have multiple local maxima and non-monotonicities. Additionally, all four corners in the empirical priors have higher probabilities than their neighborhood areas, in both the generative and preventive cases. The four corners in the probability space are associated with deterministic causal systems—if w_0 and w_1 are either 0 or 1, then the outcome is controlled deterministically based on the values of w_0, w_1, and c. This pattern found in the empirical priors is consistent with findings from previous research on human causal induction that people have a tendency to assume causal relationships to be deterministic (Frosch & Johnson-Laird, 2011; Griffiths & Tenenbaum, 2009; Schulz & Sommerville, 2006). Overall, the results based on the empirical priors seem to be in variance with the assumption that people have a sparse and strong prior.

Having empirically identified the priors on causal strength, we can now use them to predict how people reason about causes. If the empirical priors are indeed better characterizations of a population's expectations about causal systems, then a model using these priors should perform better than other models in capturing their judgments about causal strength. However, we will first turn to a discussion about how models should be evaluated. Most previous comparisons of computational models have focused on specific sets of contingencies that are designed to contrast the predictions made by different models. For example, Lober and Shanks (2000) compared the performance of the ΔP model and the causal power model by selecting sets of contingencies in which one model would predict the same value for all trials in a set while the other model would predict different values. The pattern of

Table 1
Mean and standard deviation (in parentheses) of the human judgments in the final iteration, separated by initial parameterization.

Chain	Causal direction	Small initial condition	Large initial condition
w_0	Gen.	39.95 (37.05)	59.19 (36.94)
w_0	Prev.	37.08 (36.71)	56.71 (37.06)
w_1	Gen.	80.81 (32.73)	79.26 (33.80)
w_1	Prev.	79.56 (33.48)	80.11 (33.03)

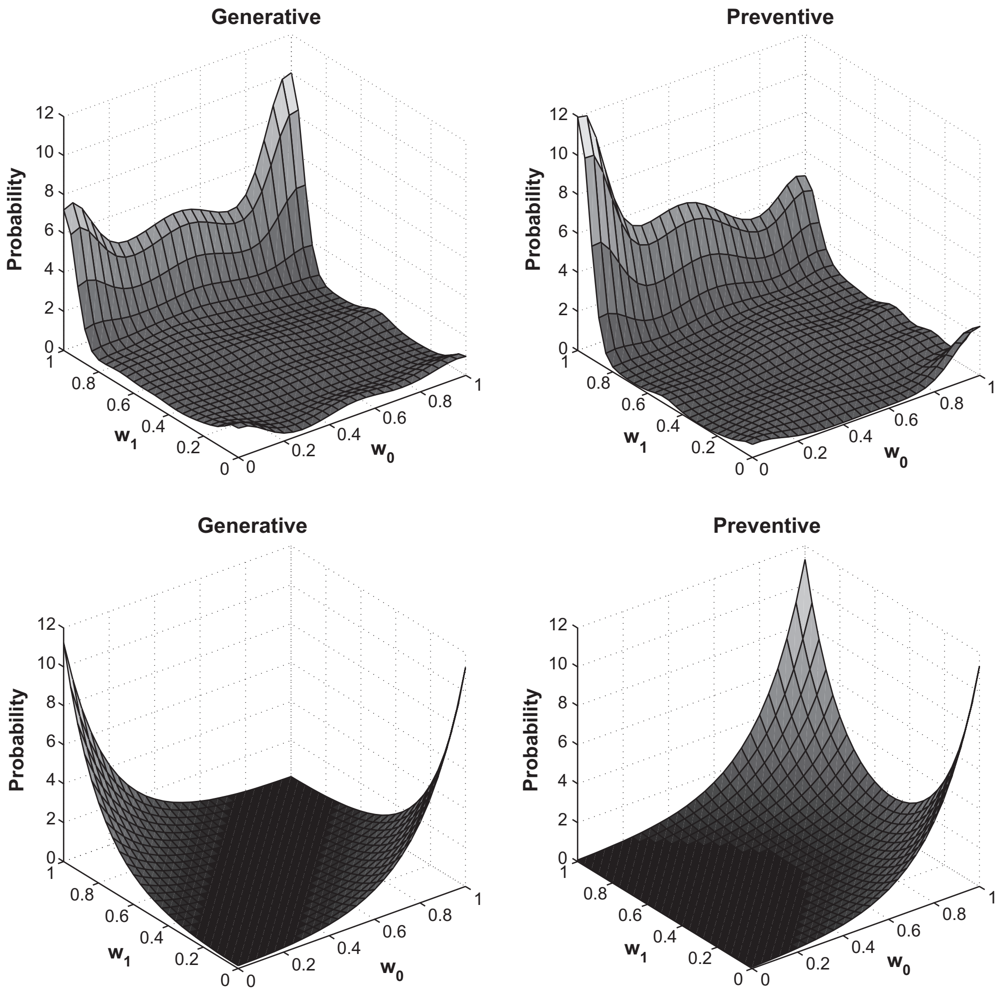


Fig. 4. Smoothed empirical estimates of human priors on causal strength produced by iterated learning in the top row. The SS priors, shown previously in Fig. 2, are reproduced here in the bottom row for easier comparison.

human judgments, especially whether they are similar or different across these trials, can then be used to lend support to either of the models under investigation. A similar approach has been used in many other studies (e.g., Allan & Jenkins, 1983; Shanks, Lopez, Darby, & Dickinson, 1996).

While using contingencies that highlight the differences between predictions by different models is a good way to understand the kinds of effects these models can accommodate, this approach creates the side effect of focusing on certain contingencies while overlooking others. This problem not only affects individual experiments and the papers based on these experiments, but also presents a challenge for meta-analyses of causal learning, as they use these papers as raw data. The coverage of contingencies in the meta-analysis conducted by Perales and Shanks (2007) provides a nice illustration of this unbalanced sampling. Their analysis included data from nine published papers, covering 19 experiments and 114 contingencies. However, as different papers had different research objectives and therefore used different contingencies to answer the respective research questions, simply aggregating these studies does not necessarily result in a balanced coverage of all possible contingencies. First, there were more contingencies for generative causes than preventive ones. If we consider both experimental instructions and ratio of contingency as cues to the causal direction, then there were twice as many generative contingencies (60) as preventive ones (30) among the 114 data points (with 20 underspecified and 4 undeterminable). More importantly, contingencies with certain characteristics were more frequently represented. In 45% of the contingencies either $P(e^+|c^+)$ or $P(e^+|c^-)$ had the value of 0 or 1 (that is, the effect was always or never observed when the cause was either present or absent), in 33% of them either $P(e^+|c^+)$ or $P(e^+|c^-)$ was equal to 0.5 (the effect was observed in half the cases either when the cause was present or absent), and in 28% of them $P(e^+|c^+) = P(e^+|c^-)$ (whether the cause was present or not did not change the percentage of effect). While the 114 contingencies overall did cover the entire contingency space reasonably well, the coverage was highly skewed towards contingencies with these specific characteristics. Perales and Shanks (2007) themselves commented on this issue, noting that "researchers tend to design and run new experiments to distinguish between theories (often only between two theories), focusing on the parts of the experimental space where their favored theory is a priori likely to be most successful" (Perales & Shanks, 2007, p. 577).

On the basis of these considerations, it seems that a broad set of stimuli that is agnostic to the researchers' hypotheses is needed in order to more comprehensively evaluate and compare the goodness of the predictions by various models. Therefore in Experiment 2 we carried out such an experiment, collecting the largest and most systematic dataset on human causal induction to date. We then used the results to compare the performance of the Bayesian models using different priors.

5. Experiment 2: A benchmark data set for causal strength judgments

Experiment 2 was designed to collect a benchmark data set for evaluating different models. To avoid the sampling bias discussed above, we conducted a large-scale study in which we collected human judgments of causal strength over the entire space of plausible contingencies, adopting an expanded version of a design previously used by Buehner et al. (2003). We aimed to include a broad representation of different contingencies by covering the entire range of values of $P(e^+|c^+)$ and $P(e^+|c^-)$ uniformly.

Another objective of Experiment 2 was to verify whether the empirical priors can be applied to contingencies of different sample sizes from those used in deriving the priors. For the empirical priors to be useful, they need to capture abstract expectations about causal relationships so that they can be applied outside of the specific setting in which they were derived. In Experiment 1, both $N(c^+)$ and $N(c^-)$ were fixed at 16. Thus in Experiment 2, we used $N(c^+)$ and $N(c^-)$ of values 8, 16, and 32, and systematically crossed them to produce more varied contingencies.

5.1. Methods

5.1.1. Participants

Similar to Experiment 1, participants were recruited from both the University of California, Berkeley subject pool, and from Amazon Mechanical Turk. Participants from the subject pool received course credit and MTurk participants received a payment of US$1. Only data from participants who completed at least 95% of the trials were included in the analysis. Subjects were randomly assigned to either the generative or the preventive condition. In each of the generative or preventive condition, there were 36 participants from the university subject pool and 180 from MTurk, for a total of 216 in each condition (and a grand total of 432). The results from the two samples were similar to each other, and were thus combined as in Experiment 1.

5.1.2. Stimuli

The stimuli used in Experiment 2 were similar to those in Experiment 1. Participants were given two samples—indicating the values of $N(e^+, c^+)$ and $N(c^+)$ in one, and $N(e^+, c^-)$ and $N(c^-)$ in the other—and were asked to make predictions about another sample. In Experiment 1, sample sizes, $N(c^+)$ and $N(c^-)$, were fixed to 16, whereas in Experiment 2, sample sizes of 8, 16, and 32 were crossed with each other, forming $3 \times 3 = 9$ *sample size conditions*. We will denote these sample size conditions by $\langle N(c^+), N(c^-) \rangle$. For example $\langle 8, 16 \rangle$ represents a set of stimuli in which $N(c^+) = 8$ and $N(c^-) = 16$.

In each sample size condition, contingencies were parameterized so that they are uniformly distributed in the contingency space. More specifically, for both the c^+ and c^- samples, we systematically varied the number of cases in which the effect is present ($N(e^+, c^+)$ and $N(e^+, c^-)$). There were nine levels of $N(e^+, c^+)$ and $N(e^+, c^-)$ in each sample size condition, with $N(e^+)$ going from 0% to 100% of the sample size in increments of 12.5%. For a sample size of 8, the nine levels included every integer from 0 to 8. For sample sizes of 16 and 32, the stimuli were in increments of 2 and 4 respectively. That is, for sample size of 16, values of $0, 2, 4, \ldots, 14, 16$ were used; for sample size of 32, values of $0, 4, 8, \ldots, 28, 32$ were used. Note that the number of cases in which the effect is absent is just the complement of the above, as $N(e^-, c^+) = N(c^+) - N(e^+, c^+)$ and $N(e^-, c^-) = N(c^-) - N(e^+, c^-)$.

Some contingencies are a priori unlikely to be observed. For example, in the generative case, a contingency of $\{\frac{N(e^+,c^+)}{N(c^+)}, \frac{N(e^+,c^-)}{N(c^-)}\} = \{\frac{3}{16}, \frac{10}{16}\}$ is very unlikely to lead to an inference of C having any causal influence because $P(e^+|c^+)$ is much smaller than $P(e^+|c^-)$. To address this, the generative condition included only contingencies in which $\frac{N(e^+,c^+)}{N(c^+)} \geqslant \frac{N(e^+,c^-)}{N(c^-)}$, i.e., $P(e^+|c^+) \geqslant P(e^+|c^-)$. Similarly, the preventive condition included only contingencies in which $\frac{N(e^+,c^-)}{N(c^-)} \geqslant \frac{N(e^+,c^+)}{N(c^+)}$, i.e., $P(e^+|c^-) \geqslant P(e^+|c^+)$.

To give a concrete example, consider the sample size condition of $\langle 8, 16\rangle$ in the generative case. $N(c^+)$ varied from $0, 1, 2, \ldots, 8$, while $N(c^-)$ varied from $0, 2, 4, \ldots, 16$. As only contingencies in which $\frac{N(e^+,c^+)}{N(c^+)} \geqslant \frac{N(e^+,c^-)}{N(c^-)}$ were included, the contingencies were $\{\frac{0}{8}, \frac{0}{16}\}, \{\frac{1}{8}, \frac{0}{16}\}, \{\frac{1}{8}, \frac{2}{16}\}, \{\frac{2}{8}, \frac{0}{16}\}, \{\frac{2}{8}, \frac{2}{16}\}, \{\frac{2}{8}, \frac{4}{16}\}, \{\frac{3}{8,\frac{0}{16}\},\ldots,\{\frac{8}{8},\frac{16}{16}\}}$. Because there were nine levels in both $N(e^+, c^+)$ and $N(e^+, c^-)$, there were $1 + 2 + \ldots + 8 + 9 = 45$ possible contingencies in each sample size condition. Multiplying that with the nine sample size conditions in each causal direction, the total number of contingencies was $45 \times 9 = 405$, in each of the generative and preventive direction.

Each participant was assigned to either the generative or preventive condition randomly. To prevent fatigue or boredom on the part of the participants from affecting experimental results, and to ensure that equal numbers of judgments were elicited from all possible contingencies, sets of 405 trials was divided among a set of nine participants in which each participant was assigned 45 trials. Within each set of nine participants, the contingencies were stratified so that each participant were given the same number of stimuli from each sample size condition, i.e., five from each of $\langle 8, 8\rangle, \langle 8, 16\rangle, \langle 8, 32\rangle, \langle 16, 8\rangle$, etc. Except for the stratification, stimulus assignment and order were randomized. As mentioned, there were 216 participants in each causal direction, resulting in $216/9 = 24$ sets of participants, and therefore 24 data points at each contingency.

5.1.3. Procedure

The procedure was the same as that of Experiment 1 with the following exceptions. To focus on the judgment of the strength of the candidate cause, we removed the question concerning the background cause. This resulted in only one prediction task in each trial. Additionally, the default location of the slider was set to 0 in the generative case and 100 in the preventive case (as opposed to having random initial positions as in Experiment 1).

5.2. Results

This experiment was designed to collect 24 human judgments on each of the 405 different contingencies from a total of 216 participants (36 from subject pool and 180 from MTurk), in each causal direction. One human judgment was missing in the generative condition due to a computer error. As a result, 9719 and 9720 human judgments were recorded in the generative and the preventive case respectively.

We computed the predictions of the three Bayesian models for each set of contingencies, and then compared them to people's judgments. Model predictions were computed by approximating the posterior distribution over a grid of uniformly spaced samples of w_0 and w_1. We first computed the prior probability at each grid point, and then multiplied them using weights proportional to the likelihood. Finally, we summed out w_0 and took the posterior mean of w_1 (Eq. (5)) as each model's prediction of human judgments (Lu et al., 2008). Except for the priors, the rest of the computations were the same

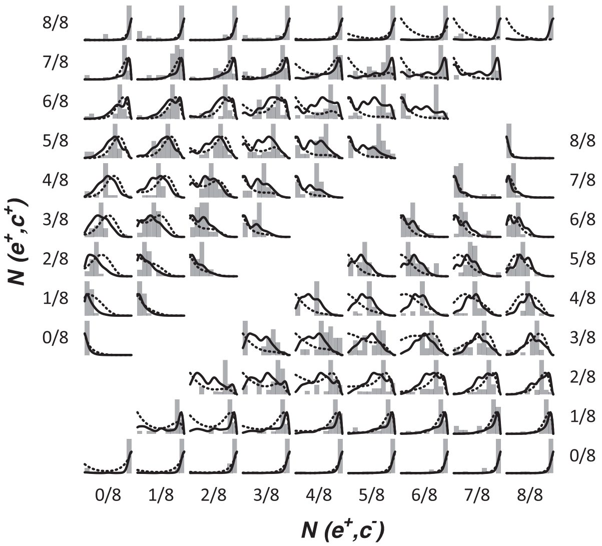


Fig. 5. Human judgments in the $\langle 8,8\rangle$ sample size condition compared against model predictions using the SS and empirical priors. The generative case is plotted in the upper left and the preventive case in the lower right. Human judgments are plotted as a histogram in gray, broken into 10 bins; predictions based on the SS and empirical priors are plotted in dotted and solid lines respectively. Contingencies that did not appear in the experiment ($P(e^+|c^+) < P(e^+|c^-)$ if generative and $P(e^+|c^+) > P(e^+|c^-)$ if preventive) are not shown.

across the three Bayesian models. Details of how these quantities were computed are provided in Appendix A.

We use one of the sample size conditions to illustrate a typical pattern of the results. Fig. 5 contrasted subjects' judgments of the strength of C against predictions by the SS and empirical models in both the generative and the preventive case at the $\langle 8,8\rangle$ sample size condition. Two properties of the Bayesian models are made apparent based on the visual inspection of this figure. First, both the SS model and the empirical model make predictions that match well with the human responses. Second, the predictions made by the two Bayesian models are not radically different, as can be seen by comparing the density curves for the SS and empirical models. This is not surprising as these Bayesian models differ only in their priors. In particular, at contingencies in which $N(e^+,c^+)$ / $N(c^+)$ is close to 0/8 or 8/8, human judgments are strongly guided by their observations and therefore the two models' predictions were quite similar. However, the models do differ in the predictions they make for cases with more mixed outcomes, in which the data did not provide strong evidence about the causal relationship. This can be seen in the middle rows (when $N(e^+,c^+)$ / $N(c^+)$ is close to 4/8) where the model predictions are more different from each other.[7]

We used four metrics to evaluate the fit between model predictions and human judgments. The metrics were Pearson's correlation coefficient (r), Spearman's rank-order correlation coefficient (ρ), root-mean-square deviations (RMSD), and a normalized version of RMSD (normalized RMSD). The correlations r and ρ compare the mean human judgments with model predictions and indicate how well they correlate with each other. These two metrics focus on the relative values of human judgments and model predictions, and do not account for the absolute difference in values. RMSD takes into account the absolute differences between judgments and predictions and is calculated as

[7] We performed a post hoc analysis comparing the performance of the empirical prior for different sample size conditions, and found that there is no significant difference.

Table 2
Comparison of model performance based on Experiment 2.

Causal direction	Metric	Uniform	SS	Empirical
Generative	r (Pearson)	0.88	0.67	0.90
Generative	ρ (Spearman)	0.90	0.67	0.92
Generative	RMSD	15.77	24.91	13.49
Generative	Norm. RMSD	12.06	17.79	10.30
Preventive	r (Pearson)	0.87	0.87	0.91
Preventive	ρ (Spearman)	0.88	0.89	0.92
Preventive	RMSD	17.62	18.34	13.49
Preventive	Norm. RMSD	16.62	17.03	12.40
Combined	RMSD	16.69	21.63	13.49
Combined	Norm. RMSD	14.34	17.41	11.35

Note: Higher values for r and ρ, and lower values for RMSD and normalized RMSD indicate better performance. The empirical model had the best performance based on all metrics.

$$\mathrm{RMSD} = \sqrt{\sum_i \frac{(x_i - \hat{x}_i)^2}{n}} \tag{8}$$

where subscript i indexes different contingencies, x_i represents the means of the human responses, $\hat{x}_i$ the model predictions, and n the number of such trials in the experiment, which is 24 in all contingencies.

While the standard RMSD weighs all trials equally, normalized RMSD adjusts for the variance of the human responses—contingencies with less variability among the responses indicate higher agreement and less uncertainty among the participants about their responses, and as a result, are weighted more heavily. The normalized RMSD is calculated as

$$\text{normalized RMSD} = \sqrt{\sum_i \frac{(x_i - \hat{x}_i)^2}{s.e._{\cdot i}} \Big/ \sum_j \frac{1}{s.e._{\cdot j}}} \tag{9}$$

where $s.e._{\cdot i}$ is the standard error of the responses at contingencies i.[8]

We analyzed the performance of the three Bayesian models using data from Experiment 2. The results are shown in Table 2. Among the three Bayesian models, all four metrics mostly agreed with each other. The correlations were the highest for the empirical model, in both the generative and preventive conditions. The results were similar for the RMSD-based metrics, with the empirical model outperforming the other two. This suggests that the empirical model does a better job of capturing the participants' expectations about the strength of causal relationships than either the uniform or SS models, in terms of both the relative and the absolute values of the judgments.[9]

To verify these results statistically, we performed an analysis using the Mann–Whitney U test to verify whether predictions made by the empirical model are indeed closer to the human responses than those made by the uniform and the SS models. We used the absolute distance between mean human judgments and the predictions by each model at each contingency as a data point (810 for both causal directions). As expected, the mean distance from the human responses for the empirical model ($M = 10.10$) was significantly lower than those of the uniform model ($M = 12.92, U = 279,729, p < 0.001$) and the SS model ($M = 15.89, U = 263,436, p < 0.001$). These results confirmed that the empirical model in fact made better predictions of the human responses than the two other models.

[8] In three of the contingencies (out of 810), all (24) participants responded with the same prediction (0 in all three cases), and therefore resulted in a zero in one of the denominators in Eq. (9) ($s.e._{\cdot j}$). In these cases, we replaced the zeroes with the next smallest possible standard error, using the standard error that would have been produced from 23 responses of 0 and one 1.

[9] The SS model incorporates a free parameter, α, that was fixed at 5 (Lu et al., 2008). We performed a grid search for the best-performing α, and found that the SS model performed best at $\alpha = 0.5$, at which value the SS prior is very similar to the uniform prior.

5.3. Discussion

In Experiment 2, we collected a large data set of human judgments in elemental causal induction, using stimuli that evenly cover all plausible contingencies. We then compared the performance of the three Bayesian models. Results showed that the empirical model, based on the priors estimated in Experiment 1 using the iterated learning technique, predicts human behavior better than other existing Bayesian models. This suggests that the empirical model represents the closest approximation of people's computations in causal induction. Furthermore, as Bayesian models rely on priors that accurately represent the reasoners' prior beliefs, the result suggests that the empirical priors best captures participants' expectations about the strength of causal relationships.

So far, both the estimation of the empirical prior and the evaluation of the resulting model used the same simple biological cover story. Having shown that the empirically estimated prior gives good predictions of causal strength judgments, a natural question to ask is how general we should expect this prior to be—whether it generalizes beyond the biological setting. We turn to this question in Experiment 3.

6. Experiment 3: Expectations about different causal systems

The empirical prior from Experiment 1 was estimated using a biological cover story concerning proteins and gene expression. As people's prior beliefs are connected to their prior knowledge about specific causal systems, it might be possible that the empirical prior estimated in Experiment 1 may also be specific to this cover story and might not correspond to people's prior beliefs about other causal systems. Therefore in Experiment 3, we varied the cover stories used in the priors estimation procedure in order to examine in what ways are people's prior beliefs concerning different causal systems similar or different.

6.1. Methods

6.1.1. Participants

Participants were 360 MTurk workers, with 90 participants in each condition. Participants received a payment of US$1.

6.1.2. Procedure

The basic experimental design was based on that of Experiment 1, with the major difference being the cover story. We used four cover stories that represent prior knowledge from four diverse domains: *physical sciences*, *medical reasoning*, *social behavior*, and *paranormal phenomenon*. Instructions in the physical condition were based on the "blicket detector" experiments that have been used to explore causal learning in children (e.g., Gopnik, Sobel, Schulz, & Glymour, 2001; Schulz, Gopnik, & Glymour, 2007; Sobel, Yoachim, Gopnik, Meltzo, & Blumenthal, 2007). In particular, we used a "super lead detector" cover story that has been used to make this paradigm more natural for adult participants (Griffiths, Sobel, Tenenbaum, & Gopnik, 2011). The instructions read:

> In this experiment, please imagine that you are working for a pencil company and you are studying the relationship between a material called "super lead" and machines called "super lead detectors". Pencil lead is made of carbon. Your company recently discovered that a new production process was resulting in a new carbon structure in their pencils—what they call "super lead". Since they are not sure which pencils they previously manufactured contain super lead, they are building a set of machines in order to detect it. These machines are programmed with different parameters to detect different types of carbon structures. You will be testing machines that are set up with different parameters.
>
> There are a number of trials in this experiment. Each trial involves a different type of super lead, and a super lead detector programmed with a different parameter set. You will see some information about how often the machine indicates the presence of super lead with a set of pencils that do not contain super lead, and how often with a set of pencils that do contain a particular type of super lead. You will be then asked to make some predictions based on these pieces of information.

In the medical condition, the cover story concerned the strength of allergy medicines in causing different types of hormonal imbalance. In the social condition, different kinds of music were played to different breeds of dogs, the inference concerned to what degree do these music cause the dogs to wag their tails. In the paranormal condition, participants read that a number of psychics used their power in trying to cause molecules to emit photons, and therefore the candidate cause and effect were the psychics' power and photon emission, respectively. Detailed instructions from the medical, social, and paranormal conditions are presented in Appendix B.

These four conditions were intended to represent causal domains about which people have different subjective beliefs. Interactions in the physical domain are often deterministic. For example, moving a magnet near objects made of iron will invariably cause an attraction force between them. Everyday life social interactions, on the other hand, are more probabilistic, since we often cannot be sure of the responses of others to our actions. The medical condition was designed to represent a causal system with elements from both the physical and the social conditions, whereas the paranormal condition was created to introduce a causal system in which participants had the least experience. We expected people's prior beliefs about these different causal systems to be reflected in the resulted prior distributions, so that we would be able to evaluate the similarities and differences among participants' prior beliefs about various causal systems. While the forms of two-dimensional probability distributions might be difficult to predict, we hypothesized that the participants hold the most deterministic expectations about the causal systems in the physical condition, followed by the medical, the social, and then the paranormal condition.

The rest of the experimental procedure was largely the same as Experiment 1. Like Experiment 1, each trial had a w_0 and a w_1 task, with a total of 48 trials—4 chains of 12 within-subjects iterations for each participant. Also similarly, the initial values of (w_0, w_1) in the four chains were $(0.3, 0.3), (0.3, 0.7), (0.7, 0.3)$, and $(0.7, 0.7)$. The iterated learning process was the same as in Experiment 1.

Part of the instructions in Experiment 3 was changed compared to Experiment 1 in order to address a potential issue. In order to elicit people's prior beliefs about a certain causal system, it is important that participants consider the entity in each trial as a different type under an umbrella category (same causal system). If the participants regarded each entity as the same type, then it is possible that their inference might rely on data from previous trials and as a result, the Markov assumption would be violated. Here we used explicit instructions to eliminate this potential issue. The instructions here emphasized that the causal relationship considered at each trial to be independent from other trials—participants were reminded in the instructions three times that both the candidate cause and effect in each trial represent different types (e.g., in each trial, a different breed of dogs listened to a different kind of music). Thus participants are reminded that while the type is different at each trial, the categories (e.g., dog in general, music in general) remain consistent. Therefore, at convergence, the hypothesis-data pair should represent the participant's prior beliefs about the causal relationship at the category level.

Beyond the cover stories and the reminders about the distinct types, there were a few differences from Experiment 1. First, we were mainly interested in how the priors associated with each cover story were similar or different with each other. Consequently the differences between the generative and preventive versions were not the main focus of this experiment, and therefore only the generative version was carried out. Second, in Experiment 3 we added three questions during the instruction phase to test the understanding of the participants about the instructions of the experiment, as advocated by Crump et al. (2013). These three questions were all multiple-choice questions and participants had to answer correctly before continuing. If participants chose the wrong answer they would be told that their answer was wrong and were allowed to try again until the right answer was selected.

6.2. Results

We estimated the prior distributions in the same manner as in Experiment 1, by pooling all participants' responses in the last iteration and then carrying out a kernel density estimation using a bivariate normal kernel. Similar to Experiment 1, the w_0 chains in the four conditions did not converge, whereas the w_1 chains in all four conditions converged quickly (by the third iteration in all conditions).

The estimated priors for the four conditions are shown in Fig. 6. Certain similarities and differences can be observed based on visual inspection. Similar to the results from Experiment 1, prior probability in all four conditions is quite high in regions associated with high w_1. This pattern is particularly extreme in the physical prior distribution. Therefore this pattern suggests that expectations of strong candidate cause might be a widely-held belief, regardless of the specific causal domain.

The medical and social priors are visually similar to each other and to the biological prior estimated in Experiment 1. This might indicate that the participants have somewhat similar prior beliefs about the causal strengths in these conditions. However, as we will see later, statistical tests will show that these two prior distributions to be distinct. While the paranormal prior is closer to the medical and the social priors than to the physical prior, it has a more pronounced high-density area at the center. This may reflect the fact that the participants were unsure about the causal relationships and responded using the middle of the scale.

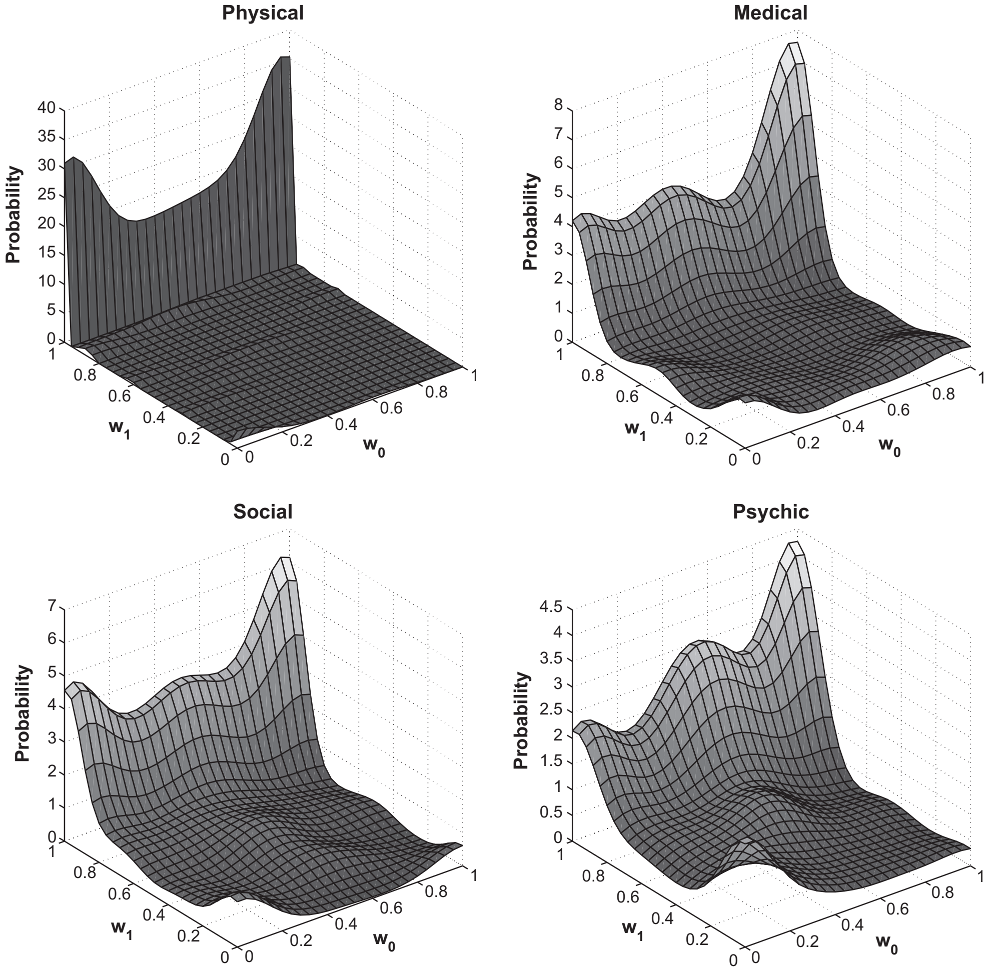


Fig. 6. Smoothed empirical estimates of the human priors for the four conditions in Experiment 3.

To formally compare the priors estimated in Experiment 3, we conducted pairwise comparisons using the *t*-test and the Kolmogorov–Smirnov test. Both tests attempt to identify differences between pairs of distributions, with the *t*-test focusing on the means and the Kolmogorov–Smirnov test focusing on non-parametric distributional differences. As there were four chains per participant and 90 participants per condition, there were 360 data points in each condition. The results are shown in Table 3, separately for w_0 and w_1. None of the differences in the tests about w_0—six *t*-tests and six Kolmogorov–Smirnov tests—were statistically significant. This suggests that the participants' expectations about the background cause were not different across different conditions. In contrast, for w_1, all tests except one (pairwise *t*-test comparison between the medical and social condition) were statistically significant, suggesting that participants' expectations about the strength of causes in these different scenarios were different.

We also compared these priors in terms of the degree of determinism, as we hypothesized that people should expect outcomes from these causal systems to be deterministic to different degrees. We expected the physical prior to represent the most deterministic beliefs, and visual inspection supports this expectation—the difference in prior probability density is most extreme in the physical prior. We then tested this formally by operationalizing the degree of determinism as the Euclidean distance between each of the responses in the last iteration and the closest corner. As the four corners—$(0,0), (0,1), (1,0)$, and $(1,1)$—represent deterministic relationships, a lower distance represents more deterministic beliefs. The lowest possible value (and highest determinism) is therefore 0, whereas the highest possible value is $\sqrt{50^2 \times 2} = 70.72$.

The degree of determinism was indeed in the order expected—the condition with the lowest distance (thus highest deterministic beliefs) was the physical condition (16.98), followed by the medical condition (19.27), then the social condition (22.23), with the paranormal condition being the highest (23.64). The differences were significant between all but the neighbors in this rank ordering (Table 3).

In all four conditions, participants favor a strong candidate cause (w_1), and this pattern is particularly strong for the physical condition. This suggests that participants expected that if the super lead

Table 3
Comparison of priors estimated in Experiment 3 based on the *t*-test, the Kolmogorov–Smirnov test, and degree of determinism.

	Medical	Social	Paranormal
***t*-test**			
w_0			
Physical	1.335 (0.18)	0.013 (0.99)	0.446 (0.66)
Medical		1.359 (0.18)	0.943 (0.35)
Social			0.445 (0.66)
w_1			
Physical	5.818 ($<$ 0.01)	5.632 ($<$ 0.01)	8.517 ($<$ 0.01)
Medical		0.546 (0.59)	2.456 (0.01)
Social			3.117 ($<$ 0.01)
Kolmogorov–Smirnov test			
w_0			
Physical	0.075 (0.26)	0.058 (0.57)	0.089 (0.12)
Medical		0.083 (0.16)	0.089 (0.12)
Social			0.047 (0.82)
w_1			
Physical	0.158 ($<$ 0.01)	0.258 ($<$ 0.01)	0.297 ($<$ 0.01)
Medical		0.111 (0.02)	0.164 ($<$ 0.01)
Social			0.108 (0.03)
Degree of determinism			
Physical	59992.0 (0.08)	54926.5 ($<$ 0.01)	52686.5 ($<$ 0.01)
Medical		59616.5 (0.06)	57403.5 ($<$ 0.01)
Social			62569.5 (0.42)

Note: The figures indicate the respective statistics with *p*-values in parentheses.

was present, it would almost always activate the super lead detector. At the same time, participants did not expect the background cause to be strong, as can be seen by the less extreme differences in density along the w_0 axis. This result can be taken as an example of how people combined a deterministic causal relationship with a probabilistic one.

Despite the differences among them, the shapes of all four priors are at variance with the idea that people have sparse and strong priors. More specifically, none of the prior distributions has peaks at both $(0, 1)$ and $(1, 0)$ as predicted by the SS theory concerning a generative cause. Furthermore, the SS theory predicts a low density area around $(1, 1)$—people do not expect that both the background and the candidate causes to be strong—but all four priors estimated in Experiment 3 have high density at $(1, 1)$. Overall, the results do not corroborate the SS theory, similar to what we found in Experiment 1.

We next computed the predictions of the Bayesian models based on these priors and evaluated their performance using the data set obtained in Experiment 2. The results are shown in Table 4. Because the human responses from Experiment 2 were elicited using the same cover story as Experiment 1, we expected the model based on the biological prior estimated in Experiment 1 to have the best performance. In contrast to this expectation, the biological model did not yield the best performance. However, as all empirical models outperformed both non-empirical models, it suggests that the general shape shared by all empirical priors—belief in a strong candidate cause—might capture an important characteristic of human causal reasoning.

To further investigate the differences in performance among the various models, we performed a bootstrapping analysis using all seven Bayesian models—uniform, SS, biological (Experiment 1), physical, medical, social, and paranormal (Experiment 3). In each bootstrap run, we sampled with replacement the participants in the Experiment 2 data set and calculated the normalized RMSD based on the sample. This procedure was repeated 1000 times for each prior. The results are shown in Fig. 7. The results support what we found previously—models using any of the five priors estimated via iterated learning outperformed Bayesian models with non-empirical priors. This suggests that the general shape of the prior distributions estimated in Experiments 1 and 3 (e.g., expectation of strong candidate cause) might capture the most important features of people's prior beliefs, with the variations across cover stories playing a less important role. We then performed a pairwise *t*-test between these seven models. All results were significant with $p < 0.001$. This is not surprising as there were 1000 simulation runs for each model, and therefore even small differences were statistically significant.

6.3. Discussion

We estimated participants' prior beliefs about different causal systems using a diverse set of cover stories. We found that while these priors can be separated quantitatively, some major features are shared between them. The most noteworthy of these features can be characterized as an expectation that the candidate cause should be strong. In spite of the diversity of the cover stories and the corresponding causal domains, none of the prior distributions estimated in Experiment 3 demonstrates features predicted by the sparse and strong theory. We also found that participants expected physical causal systems to be the most deterministic, followed by medical, social, and paranormal ones.

We compared these models in terms of how well they predict human judgments in the benchmark data set obtained in Experiment 2. Although there were differences in performance among the empirical models, they all outperformed other non-empirical Bayesian models, suggesting that the general shape of the empirical priors might be the key to understanding human causal induction.

Table 4
Performance of models based on various empirical priors estimated in Experiments 1 and 3 with respect to human data collected in Experiment 2.

Metric	Biological (Expt. 1)	Physical (Expt. 3)	Medical (Expt. 3)	Social (Expt. 3)	Paranormal (Expt. 3)
r (Pearson)	0.90	0.91	0.89	0.93	0.93
ρ (Spearman)	0.92	0.91	0.91	0.94	0.95
RMSD	13.49	13.88	15.26	10.58	11.38
Norm. RMSD	10.30	9.91	11.63	8.50	9.14

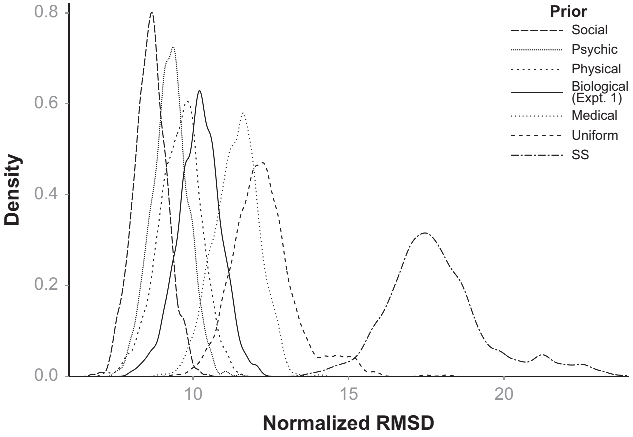


Fig. 7. Density plot of the performance of the various Bayesian models in the bootstrap analysis. Each curve represent the performance, measured in normalized RMSD, of a model based on 1000 runs. The legend is listed in the order of performance (better performance is listed first). Hence the first model listed in the legend (social) corresponds to the left-most density curve, and so on.

The family resemblance among these priors suggests that their common features might reflect a general expectation about the nature of causal relationships—a *general* causal belief that people would fall back to in the absence of strong a priori subjective beliefs concerning the specific causal system being considered. More specifically, it suggests that, in many scenarios people might expect candidate causes to be strong. Finally, these results suggest that beliefs about causal relationships might be maintained at multiple levels of abstraction—domain-general theories that prescribe the rough shape of the prior distribution and domain-specific hypotheses about particular causal systems that control the finer details.

7. General discussion

In Experiment 1, we presented an experiment using the iterated learning technique to empirically estimate people's expectations about the causal relationships in an elemental causal induction task. These expectations were operationalized as prior probability distributions over the strengths of causes. We found the resulting empirical priors to be markedly different from what have previously been proposed.

In Experiment 2, we obtained a large-scale data set that could be used as a benchmark for evaluating the performance of different models. To address the issue of selective coverage of contingencies seen in previous studies, we used a set of stimuli that uniformly covered the space of all plausible contingencies. We compared various Bayesian models of causal learning against the human responses and found that the best performance was produced by the Bayesian model using the empirical priors found in Experiment 1. This suggests that the empirical priors best capture the participants' expectations about causal relationships.

Experiment 3 used the iterated learning procedure to estimate participants' expectations about the causal relationships in a diverse set of domains. The results show that the prior distributions estimated using different cover stories share some major features (e.g., expectation of strong candidate cause) while having minor differences—suggesting that the participants have differentiated expectations while at the same time maintaining certain common theories about the nature of causal relationships. Moreover, we found that Bayesian models with empirical priors, regardless of the cover stories through which the priors were obtained, outperformed Bayesian models with non-empirical priors. This suggests that the major features captured in the empirical priors are able to explain a significant share of how people make causal induction.

In the remainder of the paper we discuss some of the issues raised by the work presented here. We will consider some of the limitations of our method, identify possible directions for future research, and highlight some of the conclusions that can be drawn from our results.

7.1. Implications for models of causal induction

The implications of our results are most direct for Bayesian models of causal induction: our analysis of models based on different prior distributions suggests that future work can be benefited from using a prior distribution similar to the empirical distributions we estimated in these experiments. Our results showed that good predictions could be obtained even using a prior distribution estimated from a different domain, so we hope that these priors will have some generality even as researchers explore different cover stories for experiments on causal induction.

This paper focuses on comparing Bayesian models with different prior distributions and has omitted comparison with non-Bayesian models (e.g., ΔP, causal power, EI rule, etc.). These non-Bayesian models do not have analogs of prior distributions, and our methodology was defined on the assumption that it was reasonable to model people's judgments in terms of Bayesian inference. However, the finding that people tend to assume causes are strong may nonetheless be useful for these other models. For example, associative models of causal learning (e.g., Shanks, 1995) need to make assumptions about the default or initial strengths of causes, and people's expectations might be relevant to setting these model parameters. Similarly, the findings here can also contribute to Bayesian models that do not use the noisy-OR/noisy-AND-NOT functional forms (Anderson, 1990; Anderson, 1991), or models that approximate Bayesian computations (Beam & Miyamoto, 2013).

The forms of the empirical priors were somewhat unexpected, as both the generative and preventive cases were significantly different from those previously proposed (Griffiths & Tenenbaum, 2005; Lu et al., 2008). However, they are consistent with previous work that suggested people have a preference for deterministic causal relationships (Frosch & Johnson-Laird, 2011; Lucas & Griffiths, 2010; Schulz & Sommerville, 2006). This finding provides a new link between formal models of causal induction from contingency data and the broader literature on causal learning.

7.2. Limitations and future directions

Our results provide a clear characterization of our participants' expectations about the strength of causal relationships, and lead to more accurate predictions of human judgments of causal strength. However, there remain a number of important issues that need to be explored in future work.

7.2.1. Origins of priors on causal strength

On identifying the structure of our participants' expectations about causal relationships, a natural question is how those expectations were formed. Following (Anderson, 1991), we might expect that people's priors reflect an adaptive response to their learning environment. In general, if people's priors capture the statistical regularities in their environment, such prior beliefs will allow them to quickly learn about the relationships (Lu et al., 2008). This adaptive proposal, however, raises many questions—how people evaluate data in their environment, how people reason about such data in their everyday lives, whether there are kinds of causal relationships that are considered more important, etc.—that we have only just begun to explore. The first step in evaluating these questions is to develop a more comprehensive picture of the contexts in which people acquire knowledge about causal relationships and their properties.

While the current paper does not attempt to answer what kind of mechanisms might have given rise to such priors, researchers have explored this issue both theoretically and empirically. For example, hierarchical Bayesian models have been suggested as a computational level explanation of how people might acquire causal knowledge (Kemp, Perfors, & Tenenbaum, 2007; Lucas & Griffiths, 2010; Tenenbaum et al., 2006). Combining advanced modeling techniques and broad data sets, such as those collected in the current paper, is likely to be the key to further developments in our understanding of the acquisition of this kind of knowledge.

7.2.2. Individual and cultural differences

An assumption of the iterated learning technique is that people share the same prior beliefs with respect to any particular causal system. While our results suggest that this is a reasonable approximation to the truth, insofar as we are able to obtain good predictions of aggregate behavior using the prior estimated by our method, it is undoubtedly a false assumption. Although labels like "people's priors" are used in this paper, the priors obtained in the current research certainly do not represent the expectations of every individual in the world. We expect that there will be differences between individuals and between cultural groups in their expectations about causal relationships. The methods that we have used here could be adapted to explore these patterns of variations.

While there has not been much cross-cultural research studying causal induction, numerous previous studies have demonstrated significant cultural differences in causal attribution (Choi, Nisbett, & Norenzayan, 1999; Morris & Peng, 1994; Nisbett, Peng, Choi, & Norenzayan, 2001). Because causal attribution is based upon causal knowledge gained previously from a lifetime of experience, it is possible that cultural differences in causal attribution might have originated from cultural differences in causal learning. Having established iterated learning to be a useful tool for investigating subjective beliefs on causal strength, we anticipate that this approach can be used to explore whether and how priors in causal learning vary across cultures.

7.2.3. Domain specificity of people's expectations

Many previous research studies on causal induction have used stimuli in the biological domain, because it provides plausible causal relationships between variables, and the functional form of these relationships is simple. Indeed, findings based on particular sets of stimuli have often been generalized broadly. However, as we demonstrated in Experiment 3, people's expectations about different causal systems can be fairly different. It is therefore prudent to reevaluate the conclusions drawn from previous findings and carefully consider the extent to which they can be generalized.

Moreover, a few points can be made with respect to interpreting the current results and integrating them with previous findings. First, the prevalence of studies with similar cover stories and formal structure means that the findings in this paper will be relevant to the ongoing research in causal induction. Second, as we have argued, the estimated prior distributions based on the several different cover stories in Experiment 3 shared a number of common features. Therefore our results and analyses provide a case study that we expect to have broad applicability. Finally, since we have developed the tools to estimate people's prior beliefs with respect to any causal systems, a similar approach can be applied in other domains to obtain a clear picture of how people make causal inferences in these situations. Indeed, the variation in the nature of the prior knowledge that informs causal induction in different domains is one of the many attractive attributes that make it a rich setting in which to explore the nature of human inductive inference (Griffiths & Tenenbaum, 2009).

7.2.4. Complexity of causal systems

An additional limitation of the current research is that, by focusing on the simplest possible causal structure, our experiments did not address how people reason about causal systems in which more than one candidate cause is present. Some recent studies have examined how people estimate the strength of multiple causes (e.g., Novick & Cheng, 2004). Adapting the iterated learning procedure to these studies might provide a way to investigate people's prior beliefs about complex causal relationships and how people reason about them.

One particularly promising application of this approach is to understand people's expectations about the functional form of causal relationships. Lucas and Griffiths (2010) recently showed that people can learn different schemes for combining the strength of causes through a relatively small amount of experience with a novel causal system. This analysis assumed a particular prior distribution on functional forms—that people favor disjunctive over conjunctive causes—and iterated learning could be used to evaluate this assumption.

7.2.5. Evaluating models of causal induction

Our results also raise questions concerning how performance of causal induction models should be measured and compared. Most previous work on computational models of causal induction compared model predictions with human responses at specific sets of contingencies. These sets of contingencies were usually chosen by researchers to highlight the different predictions made by the models being compared. We have discussed the issue concerning the distribution of contingencies used in previous studies and the need for an unbiased data set to objectively compare these models. In Experiment 2, we obtained human judgments for a wide range of stimuli that uniformly cover the space of all plausible contingencies. We see our results as providing a new benchmark against which models of causal induction can be compared. However, it could be argued that the best way to evaluate models of causal induction is to compare them based on stimuli that accurately capture the frequencies of instances of causal induction that people face in their everyday life. Developing a way to assess the natural statistics of causal relationships will an interesting direction for future research.

7.3. Conclusion

Expectations about causal strength play a key role in causal induction: they tell us what magnitude of causal influence people are looking for when they are trying to identify causal relationships. Accurately characterizing these expectations is a key problem for Bayesian models of causal induction. Since these models typically share the same assumptions about the hypotheses that people are evaluating (which are encapsulated in the causal graphical models with particular parameterizations), prior distributions are the only differentiating factor. Using iterated learning to estimate prior distributions over causal strength reveals that both the default assumption of uniform priors and the claim that people expect causes to be ‘‘sparse and strong’’ might be incorrect. Our results provide an unusually clear picture of people’s intuitions about causal relationships, and reveal that people expect causes to have a high probability of producing an effect, regardless of the rate at which that effect tends to occur.

In addition to providing this picture of people’s expectations, we have collected the largest and most comprehensive dataset of human judgments of causal strength based on contingency data. This dataset provides an important resource for future work on causal induction, providing a means of evaluating models of causal induction without concerns about biased sampling in the space of contingencies. Using this dataset, we have shown that the priors estimated through our experiments result in parameter-free Bayesian models of causal induction that outperform previous models in accounting for people’s judgments.

Beyond causal induction, we see these results as illustrating how two of the key challenges of cognitive psychology can be addressed through innovative methodologies. First, the psychological construct of interest in an area of research, such as the prior distribution on causal strengths studied in this paper, might not be quantities amenable to direct measurements. Traditionally, psychological theories and related computational models are often justified based on how well human behavior can be predicted by model output. We have presented an alternative to this approach: directly estimating the psychological construct of interest using the technique of iterated learning. Second, researchers often tend to focus on thin slices of experimental data as support for different computational models. These data might have been chosen because they demonstrate interesting phenomena, or because they allow for easier comparison between models. However, the coverage of stimuli in these studies is often quite limited. In the present paper we tackled this issue by using the larger number of participants available online to test a broad range of stimuli, in a way that is agnostic to specific hypotheses. We see this combination of new methods for measuring psychological constructs and new ways of experimental design as opening a new pathway towards resolving some of the fundamental questions of cognitive psychology.

Acknowledgments

A previous version of this work was presented at the 33rd Annual Conference of the Cognitive Science Society. We thank Doug Medin, Mike Oaksford, David Shanks, Tom Beckers, Michael Ranney, and three anonymous reviewers for helpful comments. We also thank Hongjing Lu for providing the code for the sparse and strong model, and data from the paper. This work was supported by grants FA9550-13-1-0170 and FA-9550-10-1-0232 from the Air Force Office of Scientific Research, a grant from the McDonnell Causal Collaborative, Basic Research Funds from Beijing Institute of Technology, a grant from Cher Wang and Wenchi Chen of VIA Technologies, National Natural Science Foundation of China (Program Code: 71373020), Beijing Program of Philosophy and Social Science (Program Code: 13JYB010), and the Learning Science and Educational Development Laboratory at the Institute of Education, Beijing Institute of Technology.

Appendix A. Computing causal strength estimates

We computed the predictions made by different Bayesian models used in this paper based on the procedures used by Griffiths and Tenenbaum (2005) and Lu et al. (2008). The predicted causal strength is calculated as the mean of the posterior, which is in turn based on the prior distribution of $p(w_0, w_1)$ and the likelihood $p(d|w_0, w_1)$. We will cover the computation of the likelihood and prior terms in order.

The likelihood term is shared among all Bayesian models and is based on the noisy-OR or noisy-AND-NOT parameterizations (Pearl, 2009). Given a generative cause, and w_0 and w_1, the probability of observing effect e in one case is given by the noisy-OR version:

$$p(e^+|b, c; w_0, w_1) = 1 - (1 - w_0)^b (1 - w_1)^c \tag{10}$$

in which b and c are Boolean variables representing the presence or absence of the background and the candidate cause respectively. Note that in the current model the background cause is always present. Therefore the probability of observing e with c present is $w_0 + w_1 - w_0 w_1$, whereas the probability of observing e with c absent is w_0.

For the preventive case, the probability is given by the noisy-AND-NOT version:

$$p(e^+|b, c; w_0, w_1) = w_0^b (1 - w_1)^c. \tag{11}$$

On observing contingency data d, which consists of the values of $N(e, c)$, the full likelihood in the generative case is

$$\begin{aligned} p(d|w_0, w_1) = & \binom{N(c^-)}{N(e^+, c^-)} \binom{N(c^+)}{N(e^+, c^+)} \\ & \cdot w_0^{N(e^+, c^-)} (1 - w_0)^{N(e^-, c^-)} \\ & \cdot (w_0 + w_1 - w_0 w_1)^{N(e^+, c^+)} (1 - (w_0 + w_1 - w_0 w_1))^{N(e^-, c^+)} \end{aligned} \tag{12}$$

while the full likelihood in the preventive case can be found by replacing the individual likelihood terms with ones from Eq. (11).

The prior term is formulated using the joint probability of w_0 and w_1, and is the sole difference between various Bayesian models. In the uniform model, $p(w_0, w_1) = 1$, representing the same prior probability no matter what values w_0 and w_1 take. In the SS model, the unnormalized prior probability is computed using Eqs. (6) and (7). In the empirical model, the prior probability is based on the results of Experiment 1. We took the final iteration of the human judgments and smoothed these values via kernel density estimation with a bivariate normal kernel (Venables & Ripley, 2002). The results are plotted in Fig. 4. An alternative way to compute the model predictions would be to use the human responses directly as samples from the prior, obtaining a Monte Carlo approximation to the posterior via importance sampling (e.g., Neal, 1993).

The posterior is found by multiplying the prior and the likelihood:

$$p(w_0, w_1|d) \propto p(w_0, w_1) p(d|w_0, w_1). \tag{13}$$

Posterior for either of the causal strength parameters can be found by integrating it over the value of the other. To convert this posterior into a point value, we first normalize the posterior and then take the expected value by integrating it over the desired w_i. For example, for w_1, the model prediction is:

$$\bar{w}_1 = \int_0^1 w_1 p(w_1|d) \, \mathrm{d}w_1. \tag{14}$$

In this paper the mean posteriors for all three Bayesian models were computed numerically. We first set up a grid of 51 × 51 points in the $w_0 \times w_1$ space and computed the prior density for each grid point. We then multiplied the prior with the likelihood at each grid point and normalized it to obtain the posterior. The values along the w_0 axis on this 2-dimensional grid were summed out to find the posterior for w_1. Finally the mean of the posterior w_1 was taken to be the model prediction.

Appendix B. Experiment 3 instructions

The instructions for the medical, social, and paranormal conditions were as follows:

B.1. Medical

In this experiment, please imagine that you are a researcher working for a medical company and you are studying the relationship between some allergy medicines and hormonal imbalance as a side effect of these medicines.

Your company recently discovered that a new production process was resulting in changes in the molecular structures in the allergy medicines, and these new medicines cause abnormal levels of hormones in people. Since they are not sure which medicines they previously manufactured might cause anomalies in which type of hormone, you are tasked with investigating this.

There are a number of trials in this experiment and each trial involves a different type of medicine and a different hormone. You will see some information about how often people who don't take the medicine have a particular kind of hormonal imbalance, and how often people who take that medicine have the same kind of hormonal imbalance. You will be then asked to make some predictions based on these pieces of information.

B.2. Social

In this experiment, please imagine that you are an animal researcher and you are studying the relationship between music and the tail-wagging behavior of different dog breeds.

You have found that some dogs would wag their tails after listening to some kinds of music. Since you are not sure what kind of music might cause which breed of dog to wag their tails, you have decided to investigate this.

There are a number of trials in this experiment and each trial involves a different kind of music and a different breed of dogs. For each kind of music, you will see some information about how often dogs who were not played the music wagged their tails, and how often dogs who were played the music wagged their tails. You will be then asked to make some predictions based on these pieces of information.

B.3. Paranormal

In this study, please imagine that you are a physics researcher and you are studying the relationship between psychic power and the behavior of molecules.

All molecules that you are currently investigating share a characteristic in that they all emit photons at random intervals, but at different rates. A number of psychics have claimed that they can make these molecules emit photons within a minute of when they use their power. You are tasked with investigating this.

There are a number of trials in this study and each trial involves a different psychic and a different type of molecule. For each psychic, you will see some information about how many molecules have emitted photons when a particular psychic was simply standing next to the molecules, and how many of them have emitted photons following when psychic used his/her power. You will be then asked to make some predictions based on these pieces of information.

References

Ahn, W.-k., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*(3), 299–352.

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation, 14*(4), 381–405.

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

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409–429.

Beam, C. S., & Miyamoto, J. M. (2013). *A quasi-Bayesian cascaded inference model of causal power*. Poster session presented at the annual meeting of the Psychonomic Society, Toronto, ON, Canada.

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*(6), 1119–1140.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences, 7*(1), 19–22.

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

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin, 125*(1), 47–63.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's mechanical turk as a tool for experimental behavioral research. *PloS ONE, 8*(3), e57410.

Dennis, M. J., & Ahn, W.-K. (2001). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. *Memory & Cognition, 29*(1), 152–164.

Edwards, A., Elwyn, G., & Mulley, A. (2002). Explaining risks: Turning numerical data into meaningful pictures. *BMJ, 324*(7341), 827–830.

Frosch, C. A., & Johnson-Laird, P. N. (2011). Is everyday causation deterministic or probabilistic? *Acta Psychologica, 137*, 280–291.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Suffolk, UK: Chapman & Hall/CRC.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*(5), 620–629.

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

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

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science, 35*(8), 1407–1455.

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

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

Jaynes, E. T., & Bretthorst, G. L. (2003). *Probability theory: The logic of science*. West Nyack, NY: Cambridge University Press.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review, 14*(2), 288–294.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science, 10*(3), 307–321.

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

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science, 34*(1), 113–147.

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

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: WH Freeman.

Mason, W., & Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods, 44*(1), 1–23.

Morris, M. W., & Peng, K. (1994). Culture and cause: American and Chinese attributions for social and physical events. *Journal of Personality and Social Psychology, 67*(6), 949–971.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Technical Report CRG-TR-93-1). Toronto, ON: University of Toronto.

Newsome, G. L. (2003). The debate between current versions of covariation and mechanism approaches to causal inference. *Philosophical Psychology, 16*(1), 87–107.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*(2), 291–310.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111*(2), 455–485.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). New York, NY: Cambridge University Press.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review, 14*(4), 577–596.

Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on human factors in computing systems* (pp. 2863–2872). New York, NY: ACM.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science, 10*(3), 322–332.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and children's inferences about unobserved causes. *Child Development, 77*(2), 427–442.

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge, MA: Cambridge University Press.

Shanks, D. R., Lopez, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. *Psychology of Learning and Motivation, 34*, 265–311.

Sobel, D. M., Yoachim, C. M., Gopnik, A., Meltzo, A. N., & Blumenthal, E. J. (2007). The blicket within: Preschoolers' inferences about insides and causes. *Journal of Cognition and Development, 8*(2), 159–182.

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

Sprouse, J. (2011). A validation of Amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods, 43*(1), 1–13.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*(7), 309–318.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279–1285.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. New York, NY: Springer-Verlag.

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

Williams, D. A., & Docking, G. L. (1995). Associative and normative accounts of negative transfer. *The Quarterly Journal of Experimental Psychology, 48*(4), 976–998.

Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences, 280*(1758).

Xu, J., & Griffiths, T. L. (2010). A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology, 60*(2), 107–126.