Deconfounding Hypothesis Generation and Evaluation in Bayesian Models

Elizabeth Baraff Bonawitz (liz_b@berkeley.edu)
Department of Psychology, 5427 Tolman Hall
Berkeley, CA 94720 USA

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, 3210 Tolman Hall
Berkeley, CA 94720 USA

Abstract

Bayesian models of cognition are typically used to describe human learning and inference at the computational level, identifying which hypotheses people should select to explain observed data given a particular set of inductive biases. However, such an analysis can be consistent with human behavior even if people are not actually carrying out exact Bayesian inference. We analyze a simple algorithm by which people might be approximating Bayesian inference, in which a limited set of hypotheses are generated and then evaluated using Bayes' rule. Our mathematical results indicate that a purely computational-level analysis of learners using this algorithm would confound the distinct processes of hypothesis generation and hypothesis evaluation. We use a causal learning experiment to establish empirically that the processes of generation and evaluation can be distinguished in human learners, demonstrating the importance of recognizing this distinction when interpreting Bayesian models.

Keywords: Approximate Bayesian Inference; Hypothesis Generation; Hypothesis Evaluation; Causal Learning

Introduction

Learning causal relationships, categories, and languages all require solving challenging inductive problems, using limited data to assess underdetermined hypotheses. In the last decade an increasing number of papers have argued that people solving inductive problems act in ways that are consistent with optimal Bayesian inference (e.g., Griffiths & Tenenbaum, 2005; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Xu & Tenenbaum, 2007). However, most of these analyses operate at what Marr (1982) termed the *computational* level, using Bayesian inference to identify the hypotheses that an ideal learner with particular inductive biases would choose to explain the observed data. An important question for this approach is what learners are doing at the *algorithmic* level: identifying the psychological processes by which learners solve inductive problems, and understanding how these algorithms connect back to the computational level.

Connecting the algorithmic and computational levels involves two challenges: identifying algorithms that can produce behavior consistent with Bayesian inference, and determining how the assumptions of a computational-level analysis relate to the components of these algorithms. In this paper, we take up these two challenges for one class of algorithms for inductive inference. The most naïve translation of Bayesian inference into an algorithm for inductive inference would be to assume that learners implement Bayes' rule directly, having a fixed set of hypotheses and updating a probability distribution over all of those hypotheses simultaneously as data are observed. However, the assumption that learners possess all relevant hypotheses before seeing data is at odds with numerous findings suggesting that generating appropriate hypotheses can be one of the hardest parts of inductive inference (e.g., Kuhn, 1989; Klahr, Fay, & Dunbar, 1993). We thus consider the consequences of separating the processes of generating hypotheses and evaluating those hypotheses, assuming that learners perform Bayesian inference with only the set of hypotheses they generate.

To investigate this, we present a mathematical analysis of a simple algorithm in which hypothesis generation and evaluation are separated. This produces a surprising result: This algorithm results in behavior that can still be analyzed in terms of Bayesian inference, but with a prior that conflates the plausibility of a hypothesis with the ease of generating that hypothesis. This result suggests that we should be cautious when interpreting the priors of Bayesian models estimated from behavioral data. Such priors will always reflect the inductive biases of human learners – those factors that lead people to select one hypothesis over another when both are equally consistent with the data. However, human inductive biases can include components that result from processes at the algorithmic level, such as generating hypotheses.

To demonstrate the importance of taking into account algorithmic-level factors in interpreting Bayesian models, we present an experiment exploring the separability of hypothesis generation and evaluation. In the task, we conduct a causal learning experiment in which we manipulate the hypotheses that people generate: by "priming" an appropriate hypothesis, we increase the probability of people producing responses consistent with that hypothesis; however, when we employ a more standard Bayesian reasoning task, providing a set of hypotheses and asking participants to evaluate them, the effect of priming goes away. A computational-level analysis would require postulating different prior distributions in order to explain behavior on these two components of the task. However, an algorithmic-level analysis shows that this difference can be explained as the result of the separate effects of hypothesis generation and evaluation. Finally, we discuss the implications of this work for future models of human cognition and for studies of developmental changes.

Analyzing inductive inferences

Bayesian inference indicates how a rational learner should change his or her beliefs about a set of hypotheses in light

of observed data. Let h be a hypothesis belonging to a set of hypotheses $\mathcal{H}$. Assume that the learner has different degrees of belief in the truth of these hypotheses, and that these degrees of belief are reflected in a probability distribution $p(h)$, known as the *prior*. Then, the degrees of belief the learner should assign to each hypothesis after observing data d are given by the *posterior* probability distribution $p(h|d)$ obtained by applying Bayes' rule

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h' \in \mathcal{H}} p(d|h')p(h')} \quad (1)$$

where $p(d|h)$ indicates the probability of observing d if h were true, and is known as the *likelihood*.

Bayes' rule provides a computational-level theory of inductive inference, being a component of the optimal solutions to a variety of problems of reasoning under uncertainty (Anderson, 1990; Anderson & Schooler, 1991; Freeman, 1994; Geisler, Perry, Super, & Gallogly, 2001; Griffiths & Tenenbaum, 2007; Huber, Shiffrin, Lyle, & Ruys, 2001; Knill & Richards, 1996; Körding & Wolpert, 2004; Shiffrin & Steyvers, 1997; Weiss, Simonvelli, & Adelson, 2002). As an account of inductive inference, the prior $p(h)$ captures the inductive biases of the learner, indicating which hypothesis a learner will favor when multiple hypotheses are equally consistent with the observed data (ie. which hypothesis will have higher probability when multiple hypotheses have equal likelihood). This account is attractive in that it can potentially allow us to identify the inductive biases of human learners, comparing different Bayesian models to find an appropriate prior. However, as we show in the remainder of this section, one should be cautious in interpreting such a prior: Considering algorithms by which people might be making inductive inferences shows that multiple processes can be reflected in a prior estimated from behavioral data.

Inferences with a reduced hypothesis space

As a computational-level theory of inductive inference, Bayesian models make no commitments about the psychological mechanisms by which people actually learn and reason. The most naïve interpretation of experiments demonstrating that people produce behavior consistent with Bayesian inference is that people are actually computing Bayes' rule in their heads. There are many reasons why such an algorithm is implausible, not least the requirement that people have all relevant hypotheses available whenever they make an inductive inference. However, this naïve algorithm provides a good starting point for exploring the consequences of different psychological processes that could play a role in inductive inference. Here we explore the consequences of modifying one aspect of this algorithm: rather than considering all possible hypotheses in the hypothesis space, considering only a subset of these hypotheses.

Research in inductive inference and scientific reasoning has shown that hypothesis generation is a challenging component of solving inductive problems (e.g., Kuhn, 1989; Klahr et al., 1993). Hypotheses can be generated in many different ways, including detecting cues from context, recognizing similarities to previous experiences, and making analogies to other domains (e.g., Gick & Holyoak, 1980; Gentner, 2002; Nersessian, 1992; Koslowski, 1996). We will not attempt to model these processes here, but for our purposes, it is sufficient to assume that the result of all of these processes can be summarized in a single probability distribution over hypothesis spaces. Using this probability distribution, $q(\mathcal{H}^*)$, we define the *Generate-Evaluate* (GE) algorithm for Bayesian inference with a reduced hypothesis space:

Step 1: Generate Sample a reduced hypothesis space $\mathcal{H}^* \subseteq \mathcal{H}$ from the probability distribution $q(\mathcal{H}^*)$.

Step 2: Evaluate Evaluate the hypotheses in the reduced hypothesis space $\mathcal{H}^*$ by applying Bayesian inference, using a prior distribution on $\mathcal{H}^*$ proportional to the prior on the full hypothesis space $\mathcal{H}$. Using $p(h)$ to denote the prior on the full hypothesis space, as in Equation 1 we obtain the reduced posterior distribution

$$p^*(h|d) = \frac{p(d|h)p(h)}{\sum_{h' \in \mathcal{H}^*} p(d|h')p(h')} \quad (2)$$

for $h \in \mathcal{H}^*$, with all other hypotheses receiving a posterior probability of zero. Because we are only sampling a subset of hypotheses, those that are not sampled will never be considered.

Mathematical analysis

Having defined an algorithm that takes into account the process of hypothesis generation, we can now analyze the consequences of using this algorithm. We have two questions of interest. First, will a learner using the GE algorithm produce behavior that appears to be consistent with Bayesian inference? Second, how does the process of hypothesis generation influence the interpretation of the resulting Bayesian model? We can answer both of these questions for a special case of this algorithm by exploiting its relationship to a Monte Carlo method known as importance sampling.

Monte Carlo methods are a class of algorithms that are used to approximate probabilistic computations by substituting samples from a probability distribution for the distribution itself. For example, if we wanted to perform computations involving a distribution $p(x)$, we could instead substitute a set of m values $x_1, \ldots, x_m$ drawn from $p(x)$, each with weight $1/m$. Importance sampling is a Monte Carlo method that takes this one step further, substituting samples from another distribution (the *surrogate* distribution) for samples from the target distribution (for details, see Neal, 1993). Thus, if we wanted to perform computations involving $p(x)$, we would generate a set of samples $x_1, \ldots, x_m$ from the surrogate distribution $q(x)$. We can get away with doing this if we no longer assign those samples equal weights. Instead, we give each sample x_i a weight proportional to $p(x_i)/q(x_i)$. The approxi-

mation to $p(x)$ is thus

$$p^*(x) = \frac{p(x_i)/q(x_i)}{\sum_{j=1}^{m} p(x_j)/q(x_j)} \quad (3)$$

for $x_i \in \{x_1, \ldots, x_m\}$, and zero otherwise. Intuitively, the weights proportion to $p(x_i)/q(x_i)$ reflect the "importance" of each sample. If x_i is more probable under $q(x)$ than $p(x)$, it will be over-represented in the sample, and thus should receive lower weight. If x_i is more probable under $p(x)$ than $q(x)$, there will be fewer such values than there should be, and it receives higher weight to compensate. This yields an asymptotically unbiased approximation to probabilistic computations involving the target distribution, provided certain constraints are observed (for example $q(x)$ has to be greater than zero wherever $p(x)$ is greater than zero).

Importance sampling gives us the tools we need to analyze the GE algorithm. If we assume that the samples are drawn independently, with $q(\mathcal{H}^*) = \prod_{h \in \mathcal{H}^*} q(h)$, then the GE algorithm is an importance sampler for the target distribution

$$\frac{p(d|h)p(h)q(h)}{\sum_{h \in \mathcal{H}} p(d|h)p(h)q(h)} \quad (4)$$

which is the posterior distribution obtained when using a prior proportional to the product of $p(h)$, the prior on the original hypothesis space, and $q(h)$, the probability of generating that hypothesis. It is straightforward to check that this is the case: if we approximate the distribution given in Equation 4 using $q(h)$ as the surrogate distribution, then we should generate a reduced hypothesis space $\mathcal{H}^*$ by sampling from $q(h)$ and then assign each sampled hypothesis a weight proportional to $p(d|h)p(h)q(h)/q(h) = p(d|h)p(h)$. This is exactly the procedure followed in the GE algorithm, with Equation 2 being equivalent to Equation 3.[1]

This analysis answers our two questions about the GE algorithm. First, it shows that a learner using this algorithm will still produce behavior consistent with Bayesian inference, albeit with a modified prior. Second, it indicates how the process of hypothesis generation affects behavior: If we estimate a prior by assuming people are performing Bayesian inference, that prior will reflect both the a priori plausibility of hypotheses, $p(h)$, and the probability of generating those hypotheses, $q(h)$. One needs not consider $q(h)$ when hypotheses are provided to the learner to evaluate, and thus no generation is required. However, the analysis indicates that we should be careful in interpreting priors estimated using Bayesian models: if we do not take algorithmic processes into account, hypothesis generation and evaluation are confounded. This can be problematic, as processes that change the way people generate hypotheses, such as priming a particular hypothesis, will influence the distribution $q(h)$ and hence the estimated prior, without influencing the plausibility of a hypothesis $p(h)$. Critically, ignoring the algorithmic level could therefore lead to counter-intuitive results where we need to use different priors to explain behavior across contexts where all that differs is the ease in which hypotheses are generated.

Generation and evaluation in human inferences

Our analysis assumes that the spontaneous generation of hypotheses can be separated from the evaluation of a given hypothesis. Thus, if the analysis is correct, generation and evaluation should be separable components of human inductive inference. If a learner does not sample the correct hypothesis, she will never consider it and thus cannot evaluate it; however, if a hypothesis is given to her (e.g., supplied by an experimenter), she should be able to evaluate the hypothesis just as if she generated it herself. We can empirically explore whether confounding generation and evaluation is a problem for Bayesian models in practice.

Testing the assumption that generation and evaluation are separable requires finding a task that allows us to manipulate the ease of generating different hypotheses. Previous work suggests that priming of a hypothesis can help people solve complex reasoning tasks. For example, Schunn and Dunbar (1996) found that even though participants do not spontaneously make an explicit analogy between domains, knowledge from one domain can influence reasoning in the other. Encouraged by this finding, we predicted that participants should generate different samples of hypotheses if primed differently. Priming hypotheses would thus modify the probability of generating those hypotheses, $q(h)$. However, such priming should not affect the evaluation of hypotheses provded for a learner.

In order to test whether the processes of generating and evaluating hypotheses are separable, we designed a priming task and a two part causal learning experiment. Prior to the causal learning experiment, half the participants read a vignette that primed them to think about the correct causal rule; the other half of the participants were given a "neutral" vignette. In the causal learning experiment, participants were given experience with sets of blue blocks (individuated by a letter on the block) that sometimes lit up when they interacted with each other. In the first part of the causal learning experiment, as participants encountered data, they were asked to make predictions about the result of new block interactions[2], and following all evidence, participants were asked to describe the rule they had discovered that best captured the pattern of lighting/nonlighting blocks. The actual rule by which evidence was generatd was a "rank order" rule, which meant that a latent feature of "block strength" dictated which blocks could light others. The evidence was ambiguous such that the rule was not immediately obvious, but still potentially discoverable. In the second part of the causal learning experiment, participants completed a more standard task, traditionally taken as reflecting the posterior in Bayesian learning paradigms; participants were given several rules and asked

[1] Technically, we also require that $\mathcal{H}^*$ be a multiset, allowing multiple instances of the same hypothesis.

[2] The block task was inspired by a similar method used by Tenenbaum and Niyogi (2003).

to evaluate the degree to which each rule seemed plausible, given the blocks' interactions previously demonstrated in the learning phase.

Note that because the participants are required to discover the correct causal rule in the first part of the causal learning experiment, their ability to produce the correct predictions and the correct description require both steps of the GE algorithm: the subjects must generate a set of possible hypotheses and evaluate those hypotheses to discover the causal rule that best captures the observed data. In contrast, the second part of the causal learning experiment requires only evaluation, because the set of possible hypotheses is already provided for the participant. If generation is an important factor in depermining people's inferences, we should observe a difference between the two parts of the experiment, and in particular, a difference in participants' sensitivity to the prime manipulation. Specifically, if the prime affects only generation, it should only affect participant responses in the first part of the experiment: participants given a strong prime should be more likely to generate the correct hypothesis than participants who are given a weak prime, but strong prime and weak prime participants should be equally likely to correctly rate the explanations provided to them in the second part of the experiment beacuse this task only requires evaluation and does not require generation. However, if the prime affects other things, like the prior, it will affect both parts of the experiment: the strong prime participants should not only be more likely to generate the correct causal explanations in the first part of the experiment, but they should also be more likely than the weak prime participants to provide a higher rating of the provided, correct explanation in the second part of the experment.

Methods

Participants and Design Participants were 40 undergraduates from the University of California, Berkeley who participated either for pay or for course credit. Participants were randomly assigned to either a *Strong Prime* or *Neutral Prime* condition. About half the participants completed an unrelated experiment prior to completing this experiment.

Stimuli The *Strong* and *Neutral Prime* vignettes were given to participants on a single sheet of paper with instructions. The target experiment included six small (6cm $\times$ 6cm) cardboard cutouts that the participants could manipulate as they completed the task and a 12 page booklet that included instructions, descriptions of the blocks, and sections to write in answers (see Figure 1).

Procedure The procedure involved a priming stage and a two part causal learning task, we outline each in turn.

Priming: Participants were first given an "unrelated" survey, which included a vignette about teachers watching children interacting on a playground and learning about rules that governed which children would win a game. In the *Strong Prime* condition the story suggesting that the rule governing which children would win was related to the childrens

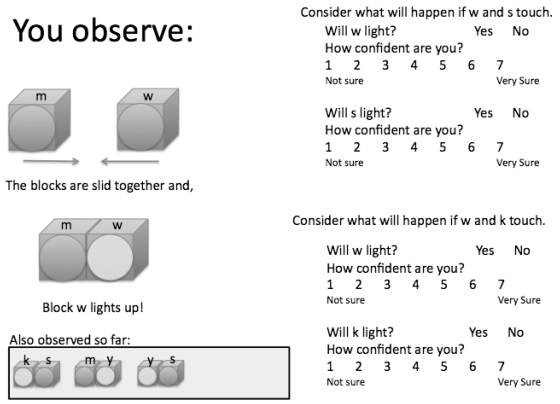

Figure 1: Example page from experiment booklet.

height. The text read, in its entirety: "Teachers at an elementary school taught their students a game for two children to play. They observed the results of pairs of students playing the game and tried to come up with a way to predict (for any given pair of students) who was going to win the game. At first it was difficult for the teachers to notice anything that would help them correctly predict the outcomes of the games. Then the teachers started organizing the children by the height of the children and the pattern of results quickly became apparent. The teachers were able to use the height of the children and make very accurate predictions as to who (for any given pair of students) was going to win the game." The *Neutral Prime* vignette was identical, except that instead of organizing the children by height, children were organized by the shirt color. Shirt color was chosen because pilot work suggested that numerous possible orderings may be plausible (e.g. sorting by the color wheel; bold colors vs. neutral colors; arranging from lightest to darkest colors, etc.), and thus the primed causal rule was somewhat arbitrary. Following the vignettes, participants were asked to respond to two simple questions about the story on the back of the sheet.

Causal Learning: In the first part of the causal learning task, participants saw sets of blue blocks (individuated by a letter on the block) that sometimes lit up when they interacted with each other. The actual rule, unbeknownst to the participants, was that the blocks could be ordered by "strength" with the "stronger" blocks always causing the "weaker" blocks to light (i.e. a variable like "height" given in the *Strong Prime* vignette, that would result in causal relations following a rank order[3]). As participants encountered data, they were asked to make predictions about the result of new block interactions ("Will this block light? Yes or no?") and provided confidence ratings on a scale of 1 to 7 (see Figure 1). Following all evidence, participants were asked to describe the rule they had discovered that best captured the pattern of light-

[3]Pilot work suggested that causal relations that follow a rank-order (e.g. dominance hierarchy) are not immediately obvious, but still potentially discoverable to participants, following suggestive evidence.

ing/nonlighting blocks and whether they could organize the blocks to best capture the rule.

In the second part of the causal learning task, participants were asked to evaluate four different explanations describing how the blocks should interact. Two explanations captured some, but not all of the data (e.g. "The blocks can be organized into two groups: blocks *s*, *k*, & *m* have the power to light up the other blocks (*y*, *w*, & *g*). Blocks in the same group do not light each other.") One explanation was nondescriptive: "The blocks can not be organized. They will light or not light randomly, but only one block can be lit at a time." And the final explanation was the target explanation, which correctly described the data: "The blocks can be organized by 'strength'. The stronger blocks will light the weaker ones. Strongest *s k m y w/g* Weakest". Participants rated the explanations on a scale from 1 ("not good") to 7 ("very good").

Results

Data were coded by the first author and reliability coded by a research assistant blind to condition and hypothesis outcomes. Explanation generation responses were labeled as "correct" or "incorrect". Agreement was 98%; the single disagreement was resolved conservatively with respect to predictions. Two participants were excluded and replaced for failing to provide a sensible response to the comprehension questions. Otherwise, all participants completed the comprehension questions for the priming vignettes.

Results confirmed that the ability to generate a hypothesis is separate from the evaluation of hypotheses. As predicted by Bayesian inference, there were no differences in evaluating the hypotheses between conditions: Both the Strong Prime and Neutral Prime participants readily rated the correct explanation equally likely: (*Strong*: 5.3; *Neutral*: 5.6; $t(38) = .48, p = ns$), and both groups ranked it well above the other (incorrect) provided rules (Strong: 2.8; Neutral: 3.0) (Wilcoxon Signed-Rank: *Strong*, $z = 3.07$, $p = .001$; *Neutral*, $z = 3.60$, $p < .001$) (Figure 2). However, there was a significant effect of condition: Participants in the *Strong Prime* condition were significantly more likely to answer the prediction questions correctly (Wilcoxon Signed-Rank: $w = 45$, $p < .01$; Figure 2a) and were more likely to generate the correct rule, Pearson $\chi^2(N = 40, 1) = 3.6, p = .058$. 65% of the participants in the *Strong Prime* condition provided the correct hypothesis, whereas only 35% of participants in the *Neutral Prime* condition generated the correct hypothesis (Figure 2b). That is, even though participants in the *Neutral Prime* condition were able to correctly evaluated the rules when they were provided, they were not necessarily able to generate the correct rule from the evidence alone.

We also looked at participant explanation ratings with the dependent factor being whether or not the participant generated the correct prediction on their own. Participants who did not generate the correct rule on their own still provided a significantly higher rating to the correct explanation (mean = 4.9) than to the incorrect explanations (mean = 3.3)(Wilcoxon Signed-Rank: $z = 2.63$, $p < .01$). That is, even though these participants were not able to generate the correct rule on their own, they were perfectly able to evaluate the good and bad explanations, being more likely to rate the correct explanation higher than the incorrect explanations.

Discussion

Connecting the computational and algorithmic levels is a significant challenge for Bayesian models of cognition. We have shown that considering the algorithms by which people might perform inductive inference can provide insight into how different psychological processes influence the conclusions that we can draw when using Bayesian models. Mathematical analysis of a simple algorithm in which learners first generate and then evaluate hypotheses indicates that while the resulting behavior is still consistent with Bayesian inference, estimation of a prior distribution from this behavior will confound the probability of generating a hypothesis and its a priori plausibility.

The responses of participants in our experiment provide some emperical support for the assumptions behind our analysis: While priming influenced whether participants could *generate* the correct explanation, it did not affect participants ability to correctly *evaluate* explanations that were provided. That is, one interpretation our our results is that the prime affected the distribution $q(h)$ from which hypotheses are generated, but it did not affect the prior probability of any particular hypothesis $p(h)$, since there were no differences between conditions when participants were asked to evaluate hypotheses that were provided to them. In the remainder of the paper, we consider some of the implications of these results and directions for future work.

Errors and approximations

Approaching inductive inference from the algorithmic level results in additional implications and predictions that may be valuable to explore in future work. For example, the algorithmic approach taken in this paper offers some reconciliation between computational level theories that suggest people are carrying out rational inference, with approaches that show people performing in seemingly "irrational" ways, such as not coming to the correct conclusion despite unambiguous or compelling evidence. By suggesting that people may be *approximating* rational inference by sampling a subset of hypotheses, these failures of inductive inference can be explained as the result of not generating appropriate hypotheses.

This makes predictions about the factors that should influence the errors that people make in inductive inference. For example, as the hypothesis space becomes large, the probability of sampling the correct hypothesis decreases. Thus, we should observe a trade-off between the size of the space and the probability of generating the correct explanation. Similarly, if cognitive limitations are imposed (for example increasing participant computational load, with additional tasks) then the set size of samples generated should decrease, and thus decrease the probability of generating the correct

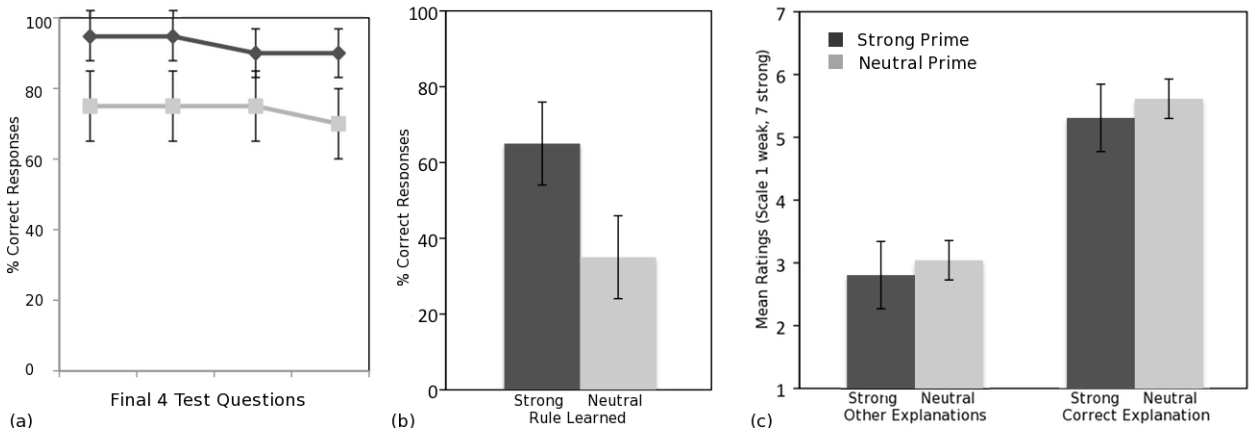

Figure 2: (a) Participants' responses to the final four prediction questions in the Neutral and Strong Prime conditions. (b) Percentage of participants who generated the correct explanation. (c) Average rating (1 weakest - 7 strongest) of the provided explanations by participants in both conditions.

sample. It may also be valuable to explore these questions in a developmental setting, examining how changes in information processing capacity influence the conclusions that children reach.

Conclusion

Bayesian models of cognition provide a computational-level account of inductive inference. Here, we have presented an anlysis that shows how taking an algorithmic-level approach can allow us to tease apart two processes that are confounded in computational-level models: hypothesis generation and evaluation. We also present experimental results that suggest that these two processes are separable in human inductive inference. Together, our analysis and emperical findings indicate that we should take both the probability of generating a hypothesis and its a priori plausibility into account when interpreting prior distributions estimated using Bayesian models. More generally, these results illustrate that understanding human inductive inference will require working at both computational and algorithmic levels of analysis, and establishing the connections between them.

Acknowledgments. We thank Nannick Bonnel and Jason Martin for assistance in data collection. This research was supported by the James S. McDonnell Foundation Causal Learning Collaborative and grant number IIS-0845410 from the National Science Foundation.

References

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

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396-408.

Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature*, *368*, 542-545.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*, 711-724.

Gentner, D. (2002). Analogy in scientific discovery: The case of johannes kepler. In L. Magnani & N. Nersessian (Eds.), *Model-based reasoning: Science, technology, values.* New York, NY: Kluwer Academic, Plenum Publisher.

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

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32:1*, 108-154.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

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

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review*, *108*, 149-182.

Klahr, D., Fay, A., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*, 111-146.

Knill, D. C., & Richards, W. A. (1996). *Perception as Bayesian inference*. Cambridge: Cambrdige University Press.

Körding, K., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244-247.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, *96*, 674-689.

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

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). University of Toronto.

Nersessian, N. (1992). How do scientists think? capturing the dynamics of conceptual change in science. In R. Giere & H. Feigle (Eds.), *Minnesota studies in the philosophy of science.* Minneapolis: Minnesota: University of Minnesota Press.

Schunn, C., & Dunbar, K. (1996). Priming, analogy, and awareness in complex reasoning. *Memory & Cognition*, *24:3*, 271-284.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review*, *4*, 145-166.

Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th annual meeting of the cognitive science society.* Hillsdale, NJ: Erlbaum.

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

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245-272.