Running Head: BAYES AND BLICKETS

Bayes and blickets:

Effects of knowledge on causal induction in children and adults

Thomas L. Griffiths

University of California, Berkeley

David M. Sobel

Brown University

Joshua B. Tenenbaum

Massachusetts Institute of Technology

Alison Gopnik

University of California, Berkeley

Address: Tom Griffiths
3210 Tolman Hall, MC 1650
University of California, Berkeley
Berkeley, CA 94720-1650
Phone: (510) 642 7134
E-mail: tom_griffiths@berkeley.edu

Abstract

People are adept at inferring novel causal relations, even from only a few observations. Prior knowledge about the probability of encountering causal relations of various types and the nature of the mechanisms relating causes and effects plays a crucial role in these inferences. We test a formal account of how this knowledge can be used and acquired, based on analyzing causal induction as Bayesian inference. Five studies explored the predictions of this account with adults and 4-year-olds, using tasks in which participants learned about the causal properties of a set of objects. The studies varied the two factors that our Bayesian approach predicted should be relevant to causal induction: the prior probability with which causal relations exist, and the assumption of a deterministic or a probabilistic relation between cause and effect. Adults' judgments (Experiments 1, 2, and 4) were in close correspondence with the quantitative predictions of the model, and children's judgments (Experiments 3 and 5) agreed qualitatively with this account.

Bayes and blickets:

Effects of knowledge on causal induction in children and adults

As adults, we know a remarkable amount about the causal structure of our environment. Discovering this structure is a difficult inductive problem, requiring unobservable causal relations to be inferred from limited observed data. Historically, psychological theories of causal induction have fallen into two camps (Newsome, 2003): covariation-based approaches characterize human causal induction as the consequence of a domain-general statistical sensitivity to covariation between cause and effect (e.g., Cheng, 1997; Shanks, 1995), while mechanism-based approaches focus on the role of prior knowledge about causal mechanisms (e.g., Ahn & Kalish, 2000; Bullock, Gelman, & Baillaergon, 1982; Shultz, 1982; Wolff, 2007). In this paper, we argue that a central part of the explanation for how people come to know so much about the causal structure of their world is that they are capable of combining these sources of information, using domain-general statistical inference guided by domain-specific prior knowledge. We show how covariational evidence and prior knowledge about causal mechanisms can be combined by a mechanism for causal reasoning based on Bayesian inference. We test the predictions of the resulting formal account through a series of experiments with both adults and children.

Bayesian inference provides a natural way to identify how covariational evidence and prior knowledge should be combined, indicating how a rational learner could best arrive at an accurate solution to the problem of inferring causal structure from observed data. The resulting computational-level analysis (in the spirit of Marr, 1982) of the problem of causal induction is analogous to work in ideal observer or signal detection theory (Green & Swets, 1966; Yuille & Kersten, 2006), which indicates how a visual system can best make inferences

about the world from visual data. Just as ideal observer models make it possible to explore how statistical information about the kinds of things encountered in the world guides perception, Bayesian inference about causal structure gives us a way to investigate how statistical information about events co-occurring interacts with existing knowledge to guide human causal learning.

In order to provide a simple, concrete setting in which to explore the interplay of covariational evidence and prior knowledge, we develop our approach for the specific case of learning about the causal relations between objects in simple physical systems. We focus on the *blicket detector* paradigm (Gopnik & Sobel, 2000; Gopnik, Sobel, Schulz, & Glymour, 2001; Sobel et al., 2004): Adults or children learn which objects (the *blickets*) have a novel hidden causal power to activate a machine (the *blicket detector*). Typically, even 4-year-olds require only a handful of observations in order to learn about the existence of this novel causal relation. Moreover, they use this knowledge both to make predictions and to design novel interventions and counterfactuals in much the same way that the causal graphical models formalism would suggest (Gopnik et al. 2004).

We use this setting to test the hypothesis that adults and children integrate prior knowledge and covariational evidence about causal relations in a way that is consistent with Bayesian inference. We explore two different kinds of prior knowledge. First, we look at the assumptions about the probability that an object is a blicket. Second, we explore a more abstract assumption about the functional form of the causal relations participants observe: whether they are deterministic or probabilistic. Our model allows us to integrate both these forms of prior knowledge with current evidence. We then examine the consequences of

modifying these assumptions through experiments in which we change the probability with which causal relations exist and whether those relations are deterministic or probabilistic.

Our strategy of conducting experiments with both adults and children illustrates the generality of our formal approach, and provides the opportunity to investigate causal induction where it is easiest to study and where it is most important. Adult participants are willing to answer a variety of questions about causality and produce multiple numerical ratings, resulting in data that are sufficiently fine-grained to allow quantitative evaluation of our models. While we can obtain only a relatively coarse characterization of the beliefs of children, they are arguably the group whose behavior we would most like to understand. Four-year-olds are still in the process of forming their deepest theories of the causal structure of their world, and using their capacity for causal induction to do so. Conducting parallel experiments with both groups provides the opportunity to test the details of our models, and to show how they might help us understand the mechanisms of cognitive development, particularly because causal graphical models explain many facets of children's causal reasoning. Further, there is a long literature on causal reasoning in young children, suggesting basic perceptual and reasoning abilities are in place at quite young ages (e.g., Bullock et al., 1982; Carey, 2009; Leslie & Keeble, 1987; Shultz, 1982), and comparing children to adults makes it possible to identify aspects of causal reasoning that might develop over time.

The Bayesian approach to causal induction that we test follows in a long tradition of formal models of human judgments about causal relations (e.g., Ward & Jenkins, 1965; Shanks, 1995; Cheng, 1997. Previous models focus on covariation between cause and effect as the basis for evaluating causal relations, and are usually applied to experiments in which such covariation is expressed over many trials on which causes and effects might occur. Our

experiments present a challenge to these models, showing that adults and children can learn causal relations from few observations, and that situations in which people observe exactly the same covariational evidence lead to different conclusions when people have different prior knowledge about the causal relations involved.

The plan of the paper is as follows. We first review formal approaches to human causal induction, and introduce the key ideas behind our Bayesian approach. We then discuss how this approach can incorporate prior knowledge on the part of the learner, and how appropriate knowledge can make it possible to learn causal relations from small amounts of data. Showing that this approach can account for some of the basic results using the blicket detector paradigm motivates our experiments. Experiments 1-3 explore the consequences of manipulating the probability that a causal relation exists in this paradigm. Experiments 4-5 examine how adults and children integrate more abstract prior knowledge with the evidence they observe by considering inferences when the mechanism between causes and effects is deterministic or probabilistic. We then consider some of the implications of these results and the limitations of our analysis in the General Discussion.

Formal models of causal induction

Most formal models of causal induction follow one of two approaches: taking causal induction to be a form of associative learning, or defining a rational procedure for estimating the strength of a causal relation. We will briefly review these two approaches.

Learning associations

One way that an individual could reason about causal relations among events in the world is to note their co-occurrence. Several psychologists have proposed that causal induction is based on recognizing associations among events (e.g., Cramer et al., 2002;

Dickinson & Shanks, 1995; Pearce, 1987). Many of these models have their origin with the Rescorla-Wagner (1972) model, which increases the strength of association between a cue and an outcome when the cue is present on a trial where the outcome occurs unexpectedly. This model provides an account of many basic findings in the associative learning literature, including Kamin's (1968) "blocking" phenomenon, in which a novel cue does not become associated with an outcome when that cue only appears in the presence of another cue that has already been associated with the outcome.

Subsequent research in associative learning has uncovered phenomena that cannot be captured easily by simple associative models. One of these phenomena is *backwards blocking*, a procedure that is identical in its contingencies to blocking, but reverses the order in which the trials are presented (Chapman, 1991; Kruschke & Blair, 2000; Miller & Matute, 1996; Shanks, 1985). In the first part of training, two cues (A and B) occur with an outcome. In the second, only one of those cues (A) occurs with the outcome. Learners come to associate only cue A with the outcome, as with the standard "forwards" blocking procedure. However, since both A and B are associated with the outcome after the first part of training, backwards blocking requires that the association between B and the outcome be modified in the absence of B in the second part of training. This is at odds with the Rescorla-Wagner model, in which associations between cues and outcomes are only modified on trials where those cues are present. A number of other phenomena that present a similar problem for simple associative models have been identified, being characterized by *retrospective revaluation* of the association between cues and effects in light of later evidence (e.g., Dickinson & Burke, 1996; Larkin, Aitken, & Dickinson, 1998).

In response to these phenomena, more sophisticated associative models have been developed in which the association between a cue and an outcome can be modified even in the absence of that cue (e.g., Dickinson & Burke, 1996; Van Hamme & Wasserman, 1994; Wasserman & Berglan, 1996). These models involve schemes for decreasing the strength of association between a cue and outcome when the outcome occurs on trials where the cue is not present. However, these models retain the basic principles of associative accounts: the inferred strength of a relation is based purely on the contingencies that hold between cues and outcomes, and the speed of learning is controlled by the extent to which outcomes are unexpected and a free parameter that sets the learning rate.

Rational methods for estimating causal strength

A second class of formal models of causal learning share with associative models the idea of estimating the strength of the relation between two variables, but rather than focusing on trial-by-trial changes, they provide an estimate that uses only the probabilities with which the effect occurs in the presence and absence of the cause. Such estimates are motivated by various rational considerations based on different construals of the nature of causal relations. Two prominent proposals in this category are the ΔP model (Shanks, 1995; Ward & Jenkins, 1965) and the Power PC model (Cheng, 1997, 2000). These models calculate an estimate of the strength of presumed causal relations given a set of data (Glymour, 2001; Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001a).

The ΔP model expresses the strength of a causal relation in terms of the efficacy of the cause in bringing about the effect. Using C to denote the cause, and E the effect, with $c+$ and $c-$ indicating occurrence and non-occurrence of the cause and $e+$ and $e-$ the corresponding values for the effect, ΔP is defined to be

$$\Delta P = p(e^+ \mid c^+) - p(e^+ \mid c^-) \quad (1)$$

This quantity has a simple interpretation in terms of the change in the probability of the effect produced by the cause, and can also be shown to be the asymptotic weight associated with the cause when the Rescorla-Wagner model is applied to contingency data generated with these probabilities (Danks, 2003).

The Power PC model (Cheng, 1997) makes further assumptions about the nature of a causal relation, essentially asserting that each cause has an independent opportunity to bring about the effect, and that its strength is the probability with which it succeeds in doing so. Under these assumptions, the strength of a causal relation (or the "causal power" of the cause) is defined by

$$\text{power} = \frac{\Delta P}{1 - p(e^+ \mid c^-)} \quad . \quad (2)$$

The denominator can be interpreted as correcting ΔP to reflect the fact that changes in the probability of the effect are more impressive when the range in which such changes can be expressed is reduced. This approach can be shown to be equivalent to assuming that causes interact via a probabilistic OR-gate (Glymour, 1998; Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001), a point that we return to later in the paper.

Prior knowledge and causal induction

Associative and rational models of strength estimation emphasize different aspects of learning, but agree in the fundamental assumptions that causal induction is a matter of determining the strength of a relation, and that covariation between cause and effect provide the information needed to solve this problem. In this section, we will argue that these models fail to capture an important part of human causal learning: the effects of prior knowledge.

Analogues of associative learning tasks have been used as the basis for experiments in causal learning with children (Gopnik et al., 2004). In one experiment, Sobel et al. (2004) presented children with a backwards blocking procedure, in which 3- and 4-year-olds were introduced to the "blicket detector", a device that lights up and plays music whenever certain objects are placed upon it. Children were shown two objects (A and B). These objects activated the detector together, demonstrated twice. Then children observed that object A activated the detector by itself. Children's inferences as to whether each object was a blicket were compared with a second condition, in which children observed two new objects (C and D) activate the detector together twice, and then object C fail to activate the detector alone. On the backwards blocking trials, children often claimed that object A was a blicket, but object B was not. Children's judgments were reversed on the other trials: object C was not a blicket, while object D was, and critically, children treated objects B and D differently even though their co-occurrence with activation of the detector was the same. Other laboratories (McCormack, Butterfill, Hoerl, & Burns, 2009) have generated similar findings for slightly older children, using other controls to ensure that children are indeed retrospectively reevaluating the probability that objects have causal efficacy.

As a result of using a backwards blocking design, these data suggest that simple models of associative learning based on the Rescorla-Wagner (1972) model may have difficulty accounting for children's causal inferences. However, this study highlights another important property of causal learning: even young children can identify a causal relation from small samples of data. Children were provided with only a handful of observations of cause and effect, yet they were confident about which objects caused the detector to activate. It is particularly constructive to compare these results with other retrospective revaluation

experiments (e.g., Kruschke & Blair, 2000; Larkin et al., 1998; Shanks, 1985), in which many more observations were required in order for participants to be similarly clear in their judgments about the underlying causal relations.

Why do we need only a handful of observations to learn a new causal relation in some situations, but dozens or even hundreds in other situations? Neither associative nor rational models of strength estimation provide a parsimonious answer to the question of what sample size will be sufficient to infer a causal relation exists. Associative models allow the learning rate to be a free parameter of the model, but this does not explain why the learning rate should be high in one situation and low in another. The rational models of strength estimation summarized above are based on the probability of the effect in the presence and absence of the cause – something that remains the same regardless of sample size. As a consequence, they draw the same conclusions from large samples and small samples, rather than requiring different sample sizes in different situations.

One way to understand why people learn quickly in some contexts but slowly in others is to appeal to prior knowledge. Consider the difference between flipping a light switch and seeing a light go on, and taking a drug and finding out that a headache goes away. When we flip the switch and the light comes on, we can be fairly sure that there is a causal relation, because we expect that if such a relation exists the light will come on every time the switch is flipped, and that there are few alternative causes that could make the light come on at that moment. We might be more reluctant to infer a causal relation when a headache goes away after taking a drug, because the relation is more probabilistic: such a drug might not work every time, and there are other factors that could cause a headache to disappear. The prior knowledge that the mechanism involved is like a light switch rather than a biochemical

process means that small samples are enough to make us confident in the existence of a causal relation. Often, however, this is not captured in existing formal models of causal induction.

Causal learning can also be fast when all that is required is a decision about whether or not a causal relation exists, rather than producing an estimate of its strength. Flipping a switch once and seeing the light come on is enough to establish a causal relation, but estimating exactly how probable it is that this will work would require much more experience. The distinction between causal structure – which relations exist – and causal strength is an important aspect of understanding causal learning (Griffiths & Tenenbaum, 2005). This distinction can give us some insight into how associative and rational strength estimation accounts might have difficulty explaining rapid learning, since both approaches focus on estimating strength as a proxy for learning causal structure.

A second interesting aspect of the backwards blocking study is that at the end of the experiment, children are sure of the status of object A as a cause, but often less certain about the status of object B. This provides another opportunity for prior knowledge to influence children's judgments. Intuitively, whether object B is likely to be a blicket should depend on how prevalent blickets seem to be: if blickets are rare, it is unlikely that B is a blicket, since the observations provide no compelling evidence to the contrary; if blickets are common, then it is more plausible that B might be a blicket.

To test whether children use this kind of prior knowledge, Sobel et al. (2004) examined whether children's judgments in a backwards blocking task were affected by the base rate of blickets. Children were first presented with a box of identical objects, and were trained that either two or ten out of twelve objects randomly selected out of the box (the *rare* and *common* conditions respectively) activated the machine. Children were then given a

backwards blocking trial, with two new objects (A and B) pulled out of the box. Children observed that A and B together made the machine go twice. Then they observed that object A made the machine go by itself. They were asked whether each of the objects was a blicket.

Across both training conditions, children identified object A as a blicket nearly 100% of the time. Information about the base rate of blickets did not affect the ability to reason about unambiguous evidence. Treatment of object B (the object not placed on the machine individually) differed between the two conditions in a way that was consistent with using prior knowledge. When blickets were rare, 4-year-olds were unlikely to categorize the B object as a blicket. When blickets were common, 4-year-olds were likely to do so.[1]

This experiment highlights another aspect of prior knowledge that informs causal induction: knowledge about the probability that a causal relation exists. When presented with ambiguous evidence, this knowledge can help guide us to a conclusion. This kind of prior knowledge is also not naturally captured within associative or rational models of strength estimation. A simple proposal might be to set the initial strength of a relation to reflect this kind of background information, but this leaves us with a new problem: how do we learn what that initial strength should be? It also does not allow us to express the difference between prior knowledge about the existence of causal relations and their strength. For example, we want to be able to differentiate between situations where 100% of causes produce the effect 10% of the time and situations where 10% of causes produce the effect 100% of the time.

These experiments suggest that at least two kinds of prior knowledge play a role in causal induction: prior knowledge about the nature of causal relations (and specifically whether they are deterministic like light switches, or probabilistic like drugs), and prior knowledge about the probability with which causal relations exist. This raises the question of

how these aspects of prior knowledge can be formalized and how they should be combined with covariational evidence in causal induction. Analyzing causal induction as a form of Bayesian inference provides a way to answer this question.

Causal induction as Bayesian inference

Bayesian inference provides a natural way to capture the effects of prior knowledge on causal induction, as it provides a rational account of how a learner should update his or her beliefs in light of evidence. A Bayesian learner begins with a space of possible hypotheses, H, where each hypothesis $h \in H$ is assigned a prior probability, $p(h)$, indicating the probability that the learner assigns to that hypothesis before seeing any data. Given observed data, d, the learner seeks to compute the posterior probability of each hypothesis, $p(h \mid d)$, indicating the degree of belief in that hypothesis in light of the data. This is done using Bayes' rule:

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')} \qquad (3)$$

where $p(d \mid h)$ is the "likelihood" – the probability of the data d under a hypothesis h, which reflects the probability distribution associated with h.

Bayes' rule allows prior knowledge to influence learning in two ways. The first, is through the prior distribution, $p(h)$, which can indicate that particular hypotheses are more likely to be true than others. For example, when hypotheses describe the existence of causal relations, certain relations could be considered more likely to exist than others. The second way in which prior knowledge can be incorporated is through the likelihood, $p(d \mid h)$. This specifies how a hypothesis relates to data, which might be different depending on the knowledge that the learner has about the process by which data are generated and the kinds of

hypotheses under consideration. In the context of causal learning, this provides a way to express different assumptions about the nature of causal relations.

To translate the general framework of Bayesian inference into a model of causal induction, we need to select hypotheses that express different possible causal relations. In this paper, the hypotheses under consideration will be specified using causal graphical models, a formal framework for representing and reasoning about causal relations (e.g., Glymour, 2001; Pearl, 1988, 2000; Spirtes, Glymour, & Scheines, 2001). These models naturally represent causal knowledge, allowing the correct kinds of inferences about interventions and counterfactuals. In addition, as probabilistic models, they also lend themselves easily to Bayesian learning methods (Friedman & Koller, 2003), and they provide a way to formalize learning causal relations that follows the assumptions behind previous psychological theories (e.g., Cheng, 1997), a point we discuss in more detail later in the paper.

A causal graphical model defines a probability distribution over a set of variables, based upon a graphical structure in which variables are nodes and edges indicate a direct causal relation (see e.g., Figure 1).

[Insert Figure 1 approximately here]

The fundamental assumption relating the underlying cause-effect graph structure to the observed probability distributions over variables is known as the *causal Markov condition*: each variable is expected to be statistically independent of its non-descendants (direct or indirect effects) given its parents (direct causes).[2] The causal Markov condition means that the joint probability distribution over the set of variables in a causal graphical model can be written as the product of a set of conditional probabilities, each indicating the probability of that variable conditioned on its parents in the graph.

Each graphical structure is consistent with an infinite set of probabilistic models that specify how the variables are related. A unique joint probability distribution \is identified by *parameterizing* the graph: defining the conditional probability distribution for each variable given its parents. Some of the simplest parameterizations assign one parameter to each edge, which can be thought of as a weight representing the strength of the corresponding causal relation. More complex parameterizations are also possible, with interactive effects that depend on the conjunctive configurations of multiple causes. The parameterization of a graph also need not be probabilistic: it is possible to specify the states of variables as deterministic functions of their causes. The form of the parameterization reflects assumptions about the nature of the mechanism by which a cause produces an effect.

Causal graphical models provide a rich language for defining hypotheses related to causal relations. They support a variety of ways in which we can define Bayesian models of causal induction. For example, we can choose to fix the causal structure and take our hypotheses to correspond to different parameter values, obtaining a Bayesian version of the rational strength estimation models mentioned above (Lu et al., 2006, 2007, 2008). Alternatively, we can take our hypotheses to correspond to different causal structures, with Bayesian inference used to determine whether a causal relation actually exists (Griffiths & Tenenbaum, 2005). In this paper, we use the latter approach – known as *structure learning* – although it may be possible to provide a similar analysis using hypotheses that vary only in causal strength. We will return to this point in the General Discussion.

Capturing constraints from prior knowledge

The Bayesian approach to structure learning makes it possible to describe formally how a learner should go about changing his or her beliefs in light of data. However, in order

to apply Bayes' rule, it is necessary to specify a hypothesis space, H, prior probabilities $p(h)$ for the relevant hypotheses, and a likelihood function $p(d \mid h)$ relating data to hypotheses. We will express the constraints that prior knowledge place on the hypotheses in two ways: by examining the ontology of objects, attributes, and event types, and through a set of causal principles relating these elements together. In this paper, we consider how to define a Bayesian model of causal learning for the kind of physical causal system instantiated in the blicket detector. The focus of this model is on the kinds of prior knowledge that inform inferences about such physical causal systems, and different models would be needed for other kinds of causal learning. Models appropriate for the more typical task in adult causal learning, with minimal prior knowledge and causes of variable strength, are presented in Griffiths and Tenenbaum (2005) and Lu, Yuille, Liljeholm, Cheng, and Holyoak (2006, 2007, 2008). A more detailed formal account of how these constraints can be expressed for a variety of causal systems appears in Griffiths and Tenenbaum (2007, 2009).

First, we define the variables that learners might consider when inferring causal relations. A simple model of the blicket detector environment might identify two kinds of entities: objects and detectors; two relevant attributes: being a blicket (a potential property of objects) and being a blicket detector (a potential property of detectors); and two kinds of events: an object being placed on a detector and a detector responding. Participants can observe the objects and events, and they are told that the machine in front of them is a blicket detector, but they must infer the remaining unseen attributes – which objects are blickets.

Our model then incorporates constraints on causal learning in the form of three assumptions, which are pieces of prior knowledge that learners might possess about blickets and blicket detectors: *temporal priority, object independence* and the *activation law*.

Temporal priority states that an object being placed on the detector causes it to activate, and that the detector's activation does not cause the experimenter to place on object upon it. Object independence holds that the identity of each object – whether it is a blicket – as well as its position in space is independent of all other objects. Object positions are visibly manipulated in the experiment but object identities are unknown, so each object is initially assigned a prior probability, ρ, of being a blicket. The activation law holds that the blicket detector activates if and only if one or more blickets are placed on top of it. This law further constrains possible causal structures: only blickets can causally influence the state of the detector. Under a strong interpretation of the activation law, this relation would be deterministic. A more realistic activation law could allow some small probability that the detector will fail to activate when a blicket is placed on it. This is a possibility that affects the predictions of this account, and one we explore in detail later in the paper, but for simplicity, we will assume for now that a detector is deterministic.

Using these assumptions, a Bayesian learner can define a constrained hypothesis space of causal graphical models for scenarios involving the blicket detector. We will initially concentrate on the backwards blocking paradigm, with two objects, A and B, and one detector.There are four hypotheses consistent with these principles (shown in Figure 1), but this set of constraints can be used to generate hypothesis spaces for arbitrary numbers of objects.[3] The activation law specifies the parameterization of those graphs, and thus the likelihood $p(d \mid h)$: E should be observed if and only if one or more of its causes is present (see Table 1). The prior probability of each hypothesis, $p(h)$, depends on the parameter ρ specified in the principle of object independence, with probabilities of $(1-\rho)^2$, $\rho(1-\rho)$, $\rho(1-\rho)$, and ρ^2 for Graphs 0-3 respectively (see the Appendix for a derivation

of these quantities, and Table 2 for a summary). For smaller values of ρ the prior probability increasingly favors causal graphs with fewer edges – as it becomes less likely that an object is a blicket, it is also less likely that there will be several blickets activating the machine.

[Insert Tables 1 and 2 Approximately Here]

The question of whether an object is a blicket can be formulated as a question of whether a causal relation exists between placing that object on the detector and the detector activating. This can be answered by computing the probability that there is an edge between an object being placed on the detector and the detector activating. For example, the probability that object A is a blicket (i.e., $A \rightarrow E$) given data d can be written

$$p(A \rightarrow E \mid d) = \sum_h p(A \rightarrow E \mid h) p(h \mid d) \qquad (4)$$

where $p(A \rightarrow E \mid h)$ is 1 if a link exists between A and E in the causal structure associated with hypothesis h, and 0 otherwise.[4]

The predictions of the Bayesian model for the backwards blocking procedure are shown in Table 2 (a detailed explanation of these predictions appears in the Appendix). According to the model, the final probability that the "blocked" object (object B) activates the machine (and hence, is a blicket), should be the prior probability of objects being blickets, ρ. If we manipulate the value of ρ, we can manipulate the magnitude of the blocking effect and the extent to which object B is judged to cause the detector to activate. Judgments about object A should be independent of ρ and equal to 1 because its efficacy has been observed unambiguously and the detector is known to be a deterministic device (via the activation law).

This simple model produces predictions that are also consistent with the results of the experiment manipulating the probability that objects are blickets in Sobel et al. (2004). Assume the prior probability, ρ, of a block being a blicket is determined in part by

observations of the base rate of blickets among objects like object B. If many objects similar to object B cause the effect, then a hypothesis in which object B causes the effect will have a relatively high initial probability. If that probability is sufficiently high, the observed data will not warrant a strong backwards blocking response, and participants will be likely to say that object B is a blicket. In contrast, if that initial probability is low, then the observed data will warrant a strong backwards blocking response, and participants will not say that object B is a blicket. This is exactly what Sobel et al. (2004) found in their experiment, and Sobel and Munro (2009) demonstrated that children only made this inference if they possess knowledge akin to the activation law we have described.

This Bayesian model explains children's ability to learn about the existence of causal relations from small amounts of data. This account predicts that manipulating prior knowledge about the probabilities with which causal relations exist (i.e. the base rate of blocks being blickets) and the nature of those relations (i.e. that the detector operates deterministically or with noise) should affect the conclusions reached by learners. Importantly, these predictions are at odds with those of associative or rational models of strength estimation. The predictions of such models depend only on the covariation between cause and effect. Our Bayesian model predicts that different conclusions can result from the same covariational evidence when different prior knowledge is used.

In the remainder of the paper, we test the predictions that result from our Bayesian model through a series of five experiments. The basic prediction is that adults and children can draw conclusions about the existence of a causal relation from relatively few observations, and that the strength of these conclusions will depend on the ambiguity of the covariational evidence and the nature of the prior knowledge to which they have access.

Experiment 1 shows that manipulating the base rate with which causal relations occur influences the extent to which adults are willing to believe a novel causal relation exists. Experiments 2 and 3 show that the strength of the conclusions that adults and children reach is influenced by ambiguity in the observed data. Experiments 4 and 5 show that the strength of these conclusions is also affected by whether causal relations are assumed to be probabilistic or deterministic.

Experiment 1: Manipulating Base Rates in Adults

Experiment 1 presented adult participants with a variant of the blicket detector backwards blocking paradigm used by Sobel et al. (2004), and asked them to make more frequent and more precise judgments of causal efficacy than children were able to answer. Working with an adult sample also enabled us to track the dynamics of learners' causal inferences. Participants rated the probability that each object was efficacious at three points during the experiment: (1) at the beginning of the trial before either object appears on the detector (the baseline rating); (2) after both objects are placed on the detector and the detector activates (the AB event); (3) after one object is seen to activate the detector by itself (the A event). These numerical judgments can be compared to the quantitative predictions of the Bayesian model (outlined in Table 1, and described in detail in the Appendix).

Method

Participants. Sixty college students were recruited from two suburban-area universities. Participants received course credit or were compensated $7/hour for their participation. Three additional participants were tested, but not included because of experimental error.

Materials. Twenty identical miniature golf pencils (approximately 3'' long) were used. The pencils were held in a large plastic cup. During the course of the experiment, Two 4" x 8" x 1" boxes were used to sort the pencils. One had the word "Super pencils" printed on it. The other had the words "Not Super pencils" printed on it.

A "super lead" detector was constructed using a desktop computer and a disguised coffee maker. Except for the 4" diameter stainless steel plate on which a coffee pot normally sits and a small black switch on the side, the coffee maker was colored exclusively white, with no markings or labels visible. The power cord emerging from the back of the coffee maker was plugged into the back of the computer; in reality, this connection was not functional, but there was no way for participants to see this. During the experiment, the apparatus was always referred to as a "super lead detector", a "super lead scanner", or a "super pencil detector". One or more golf pencils could be placed on the detector's stainless steel plate in order to "scan" them for super lead.

Participants were tested on a computer running an interface program designed to display the output of the super-lead detector. Participants observed a single window (6'' x 4.5") on the screen, with a single button (3.75" x 0.75") labeled "click to scan" and a blank frame (3.75" x 1") below the scan button that indicated the output of the detector. When the scan button was clicked with the mouse, the output frame would either light up with the words, "Super lead detected", or remain blank, depending on the precise position of the mouse arrow within the "click to scan" button. If the mouse was over the text of the words "click to scan", the detector would remain blank when the mouse was clicked; at any other position, a click would cause the words "Super lead detected!" to appear in the output box.

The experimenter always controlled the mouse, which allowed him or her to control whether the system would appear to detect super lead on any given scan.

Procedure. Participants were introduced to the experimental scenario and materials. The cup of pencils was brought out, and participants were told that some of the pencils were just normal pencils, while others were "super pencils", with a special kind of lead called "super lead". The coffee maker/computer apparatus was introduced as a "scanner" that could detect the super lead in super pencils.

The experimenter picked out one pencil at random from the cup, placed it on the detector, and clicked the "scan" button. The detector activated, with the words "Super lead detected!" appearing in the output frame of the computer screen. The experimenter said, "See, that's a super pencil", then placed the pencil in the box marked "Super pencils", and continued by saying, "Now let's try this one." The experimenter picked another pencil at random from the cup, placed it on the detector, and clicked the "scan" button. This time the detector did not activate. The experimenter said, "See, that's not a super pencil. It's just a normal one", and placed this second pencil in the box marked "Not super pencils". The two pencils were then demonstrated on the detector together, which activated. Participants were instructed that the detector activated if at least one object on it contained super lead. This demonstration paralleled Sobel et al.'s (2004) procedure. It also ensured that all participants were aware of the activation law.

Participants then took 10 pencils out of the cup, one at a time. Each pencil was placed on the detector, scanned, and then sorted into the appropriate box by the participant. Participants were assigned to one of five groups, which reflect the base rate of super pencils that they observed during the training. In group 1/6, one out of the ten pencils activated the

detector (thus 2 out of the 12 total for a base rate of 1/6). In group 1/3, three out of ten activated the detector (4 out of 12 total). In group 1/2, five out of ten did so (6 of 12 total). In group 2/3, seven out of ten did so (8 out 12 total). In group 5/6, nine out of ten did so (10 of 12 total).

After this exposure, participants were asked to take out two new pencils from the cup (referred to below as pencils A and B). They were asked to rate on a scale of 1-7 how likely each was to be a super pencil. They were instructed that a score of 1 indicated that they were confident that the object was not a super pencil, a score of 7 indicated that they were confident that the object was a super pencil, and a score of 4 indicated that they were uncertain whether it was a super pencil – that it was an even bet. After this baseline rating, the two pencils were then scanned on the machine together, and super-lead was detected (the "AB event"). Participants were again asked to rate how likely each object was a super pencil. Finally, pencil A was placed on the machine alone, scanned, and super-lead was detected (the "A event"). Participants were asked a third time to rate how likely each object was a super pencil.

Results and Discussion

Figure 2 shows the mean ratings for how likely objects A and B were to be super pencils at the three points during the experiment as well as the predictions of the Bayesian model for each of these ratings. Model predictions correspond to the posterior probabilities that causal links exist between each object (A or B) and the detector's activation (E): $p(A \rightarrow E \mid d)$ and $p(B \rightarrow E \mid d)$, respectively, computed via Equation 4 (shown above) given the hypothesis space of causal structures shown in Figure 1. We compared the Bayesian model's predictions with participants' mean ratings both quantitatively, using a measure of goodness

of fit, and qualitatively, by checking whether participants' ratings showed significant effects that were specifically and distinctively predicted by the model.

[Insert Figure 2 Approximately Here]

To assess the quantitative fit of the model, we computed the linear correlation between the model's predictions and participants' mean ratings of the probabilities of causal efficacies for 30 different judgments: each of two objects at three different points during the testing phase of all five conditions (i.e., each of the data points shown in Figure 2). The linear correlation between model predictions and human ratings over these judgments was $r = 0.97$. Note that this high correlation was achieved without the need for any free numerical parameters in the model. The only numerical parameter in the model is the prior probability ρ that an object is a super pencil, and this parameter was set equal to the base rate of super pencils that participants observed during the training phase of each condition. Hence this model fit is essentially parameter-free. For comparison, the correlation between people's judgments and the causal power of A and B, computed using Equation 2 (the Power PC theory), is $r = 0.325$.[5] This low correlation results from the fact that causal power does not take into account the variation in judgments produced by changing the base rate with which causal relations exist (see the General Discussion for details).

In addition to making quantitative predictions about the ratings provided by our participants, the Bayesian approach makes four qualitative predictions that collectively discriminate it from other accounts. The first qualitative prediction is that initial ratings (before objects A or B are placed on the detector) should reflect the prior probability of encountering a super pencil. Since prior probabilities are not used in associative or strength-based approaches, this prediction is unique to the Bayesian model. Preliminary analysis

revealed no difference between the initial ratings of the A and B pencils overall, $t(59) = 1.00$, *ns*. As a result, these ratings were averaged together. An analysis of variance showed that across the five groups, these initial ratings significantly differed, $F(4, 55) = 63.11$, $p < .001$, Partial $\eta^2 = 0.82$. Post-hoc analysis revealed that each group's ratings was significantly different from the adjacent group (group 1/6 was significantly lower than group 1/3, group 1/3 was significantly lower than group 1/2, etc.), all *p*-values < .05 with Tukey LSD correction. These results suggest that participants recognized that the base rate of super pencils differed among the five conditions – as the base rate increased, so did participants' initial ratings.

The second prediction of the Bayesian account is that after the AB event, ratings of objects A and B should increase above their baseline levels, but this increase should be smaller as the base rate of super pencils increases. Again, this prediction is unique to the Bayesian model because the other accounts do not use base rate information. The ratings between the A and B pencils at this point in the trial did not significantly differ, $t(59) = 0.91$, *ns*., so these ratings were averaged together, and compared with the average initial ratings using a 2 (Initial vs. After AB event) x 5 (Condition) mixed Analysis of Variance. This analysis revealed a main effect of event; participants ratings increased overall between the initial and AB event, $F(1, 55) = 67.87$, $p < .001$, Partial $\eta^2 = 0.55$. A main effect of condition was also found; overall, ratings differed among the five conditions, $F(4, 55) = 50.99$, $p < .001$, Partial $\eta^2 = 0.79$. Critical to the prediction, a significant interaction between condition and rating was also found, $F(4, 55) = 17.33$, $p < .001$, Partial $\eta^2 = 0.56$.

To examine this interaction, we computed difference scores between average ratings of objects A and B after the AB event and ratings of these objects at the beginning of the trial. The Bayesian model predicts that these difference scores should be higher as the base rate of

super pencils decreases. Participants' difference scores were significantly higher in the 1/6 condition than the 1/3 condition, $t(22) = 2.03$, $p = .05$, Cohen's $d = 0.83$, and were significant higher in the 1/3 than the 1/2 conditions, $t(22) = 4.84$, $p < .001$, Cohen's $d = 1.97$,. Difference scores did not differ between the 1/2 condition and either the 2/3 or 5/6 conditions, both $t(22)$-values < 1.09, both p-values *ns*. Specifically, the average ratings of the A and B pencils were not significantly different from the average initial ratings of these objects in the 1/2, 2/3 and 5/6 conditions. They were significantly different in the 1/6 and 1/3 conditions, $t(11) = -10.32$ and -5.53, respectively, both p-values < .001, both Cohen's d-values > 1.94.

The third prediction of the Bayesian account is that after object A activates the detector by itself, ratings for it should be at ceiling. This prediction is not unique to the Bayesian account – it might be possible for other accounts to make a similar prediction. For example, a similar prediction could result from associative learning with a high learning rate, computation of causal power from the contingencies of A and E, and deductive reasoning under the assumption that occurrence of E in the presence of A indicates a dependency between these variables. Nonetheless, if this prediction were inconsistent with the data, it would provide evidence against the Bayesian account. That said, the prediction that ratings for A should be at ceiling after activating the detector was borne out in the data. Across all five conditions, every participant rated object A at ceiling levels at the end of the trial (i.e., 7 out of 7).

The final prediction is that at the end of the trial, ratings of object B should decrease back to their baseline levels. Ratings did return to baseline levels; no significant differences were found between the initial ratings of the objects and the ratings of object B at the end of the trial for all conditions with one exception: In the 1/2 condition, ratings of object B were

significantly lower at the end of the trial than the initial ratings of the objects, $t(11) = -2.57$, $p = .026$, Cohen's $d = 1.05$ (all other $t(11)$-values < -1.56, all p-values *ns*.).

The present data are both quantitatively and qualitatively consistent with the predictions produced by the hypothesis that human causal reasoning can be explained as a kind of Bayesian inference, guided by appropriate domain knowledge. They are also inconsistent with standard associative and rational strength estimation models. First, they illustrate a rapidity of causal learning that is strikingly different from that seen in other backwards blocking studies (e.g. Kruschke & Blair, 2000; Larkin et al., 1998; Shanks, 1985). Rational strength estimation models are insensitive to sample size, and while associative models could fit the resulting data by changing the learning rate, they provide no explanation for why the learning rate should be different between these two settings, while the Bayesian approach naturally explains this difference in terms of appropriate prior knowledge. Second, our results show that people's judgments are sensitive to base rates, while both associative and rational strength estimation models assume that only covariational evidence is used in evaluating causal relations.

We provide a more detailed comparison to the predictions of alternative accounts in the General Discussion, but one alternative is sufficiently compelling that we will consider it here. This is the possibility that our participants might not be using base-rate information to establish a prior probability and then integrating it with later evidence in a Bayesian way. Instead, they might simply use base rate information as a heuristic after they have determined that there is not enough evidence to make a deductively valid inference (a strategy consistent with previous work emphasizing the deductive component of causal reasoning, e.g., Johnson-

Laird, 1994; Goldvarg & Johnson-Laird, 2001). We refer to this heuristic style of causal reasoning as "deduction, with a default to base rate," or DBR for short.

More precisely, the DBR heuristic would treat the backwards blocking sequence as follows: two objects are brought out and together they activate the machine. Since participants were trained that individual objects labeled "super pencils" activate the detector, there are one of three possible manners of resolving this event: either object A is the only super pencil, object B is the only super pencil, or they are both super pencils. Object A is then demonstrated to unambiguously activate the machine, so it is definitely a super pencil. Deductive reasoning now indicates that there is no relevant evidence for object B. In this case, participants might explicitly revert to the base rate to make a probability judgment about object B as a simple fallback heuristic rather than as the consequence of rational Bayesian updating. If super pencils are common they judge that it is a super pencil, and if super pencils are rare they judge that it is not. Although we do not know of experimental support for this account, this approach seems consistent with modifications of deductive inference accounts of blocking phenomena suggested by Lovibond and colleagues (e.g., Lovibond, Been, Mitchell, Bouton, & Frohardt, 2003; Mitchell, Killedar, & Lovibond, 2005; see also McCormick et al., 2009).

In Experiments 2 and 3, we contrasted the predictions of this account of causal reasoning with the Bayesian approach by presenting learners with only ambiguous data. Adults and 4-year-olds, respectively, were trained that super pencils or blickets were infrequent in the same manner as in the 1/6 condition. They were then shown evidence in which no single object was ever placed on the detector alone. DBR reasoning would not lead to correct judgments in this case, because no unambiguous data are presented to support a

deductive inference. These experiments also provide us with a further opportunity to explore rapid causal learning, and to show how ambiguity in covariational evidence combines with prior knowledge to determine how people evaluate a novel causal relation.

Experiment 2: Learning from Ambiguous Evidence in Adults

In Experiment 2, a new set of adult participants were trained that "super pencils" were rare and then shown three objects (A, B, and C). Objects A and B activated the machine together. Then objects A and C activated the machine together. Participants were asked to rate their belief that each object was a "super pencil" at three points during the trial: when they were first brought out, after objects A and B activated the detector together, and after objects A and C activated the detector together. This design ensured that participants received only ambiguous evidence concerning the properties of each of the three objects, and follows in a tradition of similar procedures that have been conducted in the associative learning literature (e.g., Cole, Barnet, & Miller, 1995).

With three objects that can potentially activate the detector, the hypothesis space defined by the principles underlying our Bayesian model contains eight causal structures, shown in Figure 3 The prior and posterior probability distributions over these hypotheses after the AB and AC events are shown in Table 3. The explanation for these distributions is similar to that given for backwards blocking in the Appendix. The AB event rules out Graph 0 and Graph 3, and the AC event rules out Graph 2, but many candidate causal structures are consistent with the data: A alone is a super pencil (Graph 1), A and B are super pencils (Graph 4), A and C are super pencils (Graph 5), B and C are super pencils (Graph 6), or all three objects are super pencils (Graph 7). The actual posterior probabilities of these structures depend upon the prior. If super pencils are rare (i.e., ρ is low), then structures with fewer

causal links will have higher initial prior probabilities, and ultimately higher posterior probabilities.

[Insert Table 3 and Figure 3 Approximately Here]

The posterior probabilities shown in Table 3 express a set of qualitative predictions. When asked to judge the probability that object A is a super pencil, participants should consider the implications of all candidate causal structures consistent with the data, weighted by their relative probabilities, in accord with Equation 2. The only inconsistent models (with a posterior probability of zero) are the ones in which there are no super pencils, or where object B or C is the only super pencils (Graphs 0, 2, and 3) and so those models should be eliminated.

Looking across the remaining models, the $A \rightarrow E$ edge occurs in all but one consistent model (i.e., Graph 6). The $B \rightarrow E$ and $C \rightarrow E$ edges are present in fewer consistent models, but they do occur, so they also accrue some posterior probability value. Thus, at the end of the trial, object A should be judged most likely to be a super pencil, but not at ceiling values, since a model like Graph 6 has non-zero posterior probability. The probability that Objects B and C are super-pencils should be higher than the base rate, but lower than the probability that A is a super-pencil. This is because B and C are super-pencils in some of the possible models.

A critical difference between the present experiment and Experiment 1 is that in Experiment 1, all models in which object A was not a super pencil had a posterior probability of zero, and hence the Bayesian model predicted ceiling performance. Here, that is not the case, and the model specifically predicts below-ceiling performance. Similarly, in Experiment 1 the model predicts that ratings for B at the end of the trial will fall back to the base rate. In this experiment, ratings for B and C at the end of the trial should be higher than the base rate.

Critically, these predictions hold most strongly for a low base rate (i.e., ρ is low). When ρ is high, we would still expect change in adults' probability judgments in the same pattern, but not to the same extent (because their baseline ratings would be higher). We thus established a context in which superpencils are rare in order to maximize the strength of the effect.

Method

Participants. Twenty-one college students were recruited from a suburban-area university's psychology subject pool, with the same demographics as in Experiment 1. One participant was excluded for failure to understand the experimental instructions (see below), leaving a sample of twenty. Participants received course credit for their participation.

Materials. The same "super pencil" display and set of golf pencils as in Experiment 1 were used.

Procedure. All participants were introduced to the super pencil detector in the same manner as Experiment 1. All participants were given the same training as participants in the 1/6 condition of Experiment 1. Thus, all participants observed that only two out of twelve pencils, chosen at random, activated the detector.

In the test phase of this experiment, participants chose three pencils out of the box. The experimenter labeled the pencils "A", "B", and "C", in arbitrary order, to keep track of individual objects through the remainder of the study. Participants were first asked to rate how likely they thought each of the three objects was to be a super pencil. Two objects (e.g., A and B) were placed on the detector, which activated (the AB event). Participants were again asked to rate how likely each object was to be a super pencil. One of the pencils that had just activated the machine (e.g., A) and the third pencil (e.g., C) were placed on the detector together, which again activated (the AC event). Participants were again asked to rate how

likely each object was to be a super pencil. The ratings were always on a scale of 0 to 10, with 0 indicating that the object is definitely not a super pencil, 10 indicating that it definitely is, and 5 an even bet. The spatial configuration and arbitrary labels of the three objects were counterbalanced across participants. For convenience, in this paper we will use the canonical labeling of objects given above, even though different participants observed different objects in each role. (For instance, some participants saw the "B" pencil or the "C" pencil placed on the detector twice, in a counterbalanced fashion).

Finally, participants were given a debriefing form in which they were asked to describe how they made their judgments. One participant stated that her judgments were made at random, based on the apparent sharpness of each pencil, and her data were excluded from the analysis. All other participants referred to some form of intuitive statistical reasoning, though they were unable to provide much detail.

Results and Discussion

Preliminary analysis revealed no effect of spatial location on ratings at any point. There was also no difference between ratings of the three objects at baseline, so these were averaged. Preliminary analysis also revealed no difference between the ratings of objects A and B after the AB trial or between objects B and C at the end of the procedure, so these data were averaged. Figure 4 shows participants ratings of the three objects at the various stages across the AB-AC sequence, and the predictions of the Bayesian model, calculated in the same manner as in Experiment 1.

[Insert Figure 4 Approximately Here]

As in Experiment 1, model predictions correspond to the posterior probabilities that causal links exist between each object (A, B, or C) and the detector's activation (E): $p(A \rightarrow E \mid$

d), $p(B \rightarrow E \mid d)$, and $p(C \rightarrow E \mid d)$, respectively, computed via Equation 4 and using the hypothesis space of causal structures shown in Figure 3. We again compared the Bayesian model's predictions with participants' mean ratings both quantitatively, using a linear correlation measure, and qualitatively, by checking whether participants' ratings show significant evidence for the effects predicted above.

Linear correlation between the model's predictions and people's mean ratings of the probabilities of causal efficacies was computed for nine different judgments: each of three objects at three different points during the testing phase of the experiment (i.e., each of the data points shown in Figure 4). The linear correlation between model predictions and human ratings over these nine judgments was $r = 0.98$. Again, this high correlation was achieved without setting any free numerical parameters in the model, because the prior probability ρ that an object is a super pencil could be set equal to the base rate of super pencils that participants observed during the training phase of the experiment. For comparison, the correlation between people's judgments and the causal power of A and B, computed using Equation 2, is $r = 0.768$ (see the General Discussion for details).

The average rating of the three objects as super pencils before any of them were presented on the detector was 2.48 on a scale of 0 to 10. This was significantly less than the average rating of objects A and B after they were placed on the detector together, in the AB trial (4.60), $t(19) = -5.36$, $p < .001$, Cohen's $d = 1.34$.[6] After the AC trial, in which objects A and C are placed on the detector, the average rating of object A was 6.70. This rating was significantly below ceiling level (10), one-sample $t(19) = -6.13$, $p < .001$, and was greater than the average rating of objects B and C (3.43), which each had been placed on and activated the detector once, $t(19) = 4.83$, $p < .001$, Cohen's $d = 1.53$. This rating was significantly lower

than ratings for objects A and B after the AB event, 3.43 vs. 4.60, $t(19) = 2.67$, $p < .05$, Cohen's $d = 0.70$. These differences are all predicted by the Bayesian model (see Table 3).

We showed that when adult learners were given only ambiguous data, they could integrate probabilistic evidence and prior knowledge in an approximately Bayesian fashion to infer unobservable causal relations with appropriate degrees of confidence.

The four levels of response, all greater than floor value and lower than ceiling value, are predicted by the Bayesian model, and separate it from alternative accounts of causal inference. Further, these findings cast doubt on simpler heuristic approaches, such as the DBR heuristic, in which causal reasoning is based on deductive logic and prior probabilities are used only as a last resort.

Experiment 3: Learning from Ambiguous Evidence in Children

Although previous research (Sobel et al., 2004) found that 4-year-olds are sensitive to a base rate manipulation using the backwards blocking paradigm from Experiment 1, it is unclear whether children have developed a Bayesian mechanism for causal inference. Children might adopt some other approach to causal learning, such as the DBR heuristic described in Experiment 1. Experiment 3 tested 4-year-olds on a version of the AB-AC paradigm from Experiment 2.

A complication comes from the fact that we cannot expect children to make stable fine-grained numerical judgments of subjective probability as adults did. Our response measure was the same yes/no question ("Is it a blicket?") used in previous developmental studies. We measured whether children's judgments were qualitatively similar to those of adults. We assumed that the number of times children said that an object was a blicket reflected their subjective probability assessment of whether it was a blicket. This also meant

that we could not assess the child's judgments at each step of the trial. To deal with this problem, after the AB-AC sequence, we gave children another trial, in which they saw two new objects (X and Y) activate the detector together, and were asked to categorize each of them. This trial was similar to making a rating about just the intermediate AB event in Experiment 2 (because children could only use the evidence of the base-rate and the effects of X and Y together). As in Experiment 2, we kept the base rate of objects having causal efficacy low.

Method

Participants. The final sample was composed of sixteen 4-year-olds (6 girls, M = 54.06 months, range 49-58 months), recruited from a suburban area university preschool and from a list of hospital births provided by an urban area university. Two additional children were tested, but excluded for failing control questions (see below).

Materials. The "blicket detector" used by Sobel et al. (2004) was used here. The detector was 5" x 7" x 3", made of wood (painted gray) with a red lucite top. It "activated" as soon as particular objects (controlled by the experimenter) made contact with it and continued to light up and play music as long as an object made contact with it. This provided a strong impression that something about the object itself caused the effect.

Eighteen blue wooden cylindrical blocks were used. These blocks were held in a 12" x 12" x 4" white cardboard box. Two smaller 6" x 12" x 2" white cardboard boxes were also used. One had the word "Blickets" printed on it. The other had the words "Not Blickets" printed on it. Two white metal knobs (approximately 1½" in diameter) and two small metal tee-joints (approximately 1½" in length) were also used.

Procedure. Children were tested by an experimenter with whom they were familiar. Children were first given a pretest. The two metal knobs and two tee-joints were placed in front of the child. Children were told that one of the metal knobs was a "dax" and were asked to give the experimenter the other dax. After they responded, children were told that one of the metallic tee-joints was a "wug" and were asked to give the experimenter the other wug. The pretest ensured that children would extend novel names onto objects and would interact with the experimenter. Children were then shown the blicket detector. They were told that it was a "blicket machine" and that "blickets made the machine go." The box of blocks was brought out and children were told that blickets were infrequent: "I have this whole box of toys and I want to know which ones are blickets. It's a good thing we have this machine because only a few of these are blickets. Most of these are not. It's very important to know which are which." Two blocks were then taken out of the box and the experimenter said, "Let's try these two." The blocks were placed on the machine together and the machine activated. The experimenter said, "Look, together they make it go. Now let's try them one at a time." One of the two blocks was then placed on the machine and the machine activated. The experimenter said, "Wow. Look, this one makes the machine go by itself. It's a blicket. I have this box and it says 'blickets' on it. Let's put the blicket in the blicket box." The experimenter put the block that just activated the machine into the white box labeled "blickets". The experimenter then said, "Now let's try this other one." The other object was put on the machine and it did not respond. The experimenter said, "Wow. Look, it did not make the machine go by itself. It is not a blicket. I have another box that says 'Not Blickets' on it. Let's put this one in the 'Not blicket' box."

Next, the experimenter said, “Remember, when we did them together – together they made the machine go.” This was demonstrated with the two blocks. “But this is because the blicket made it go and the not blicket didn’t make it go.” Each block was demonstrated individually with its proper effect on the machine. This provided children with information about the activation law: to demonstrate that even if only one block on the machine was a “blicket,” the machine would activate.

Five pairs of blocks were taken out of the box and each was demonstrated on the machine. Only one out of the ten blocks made the machine go (randomly determined). After each pair was demonstrated, children were asked which box each object went into. After the child made their response, the experimenter confirmed it by asking, “Just to make sure, is this one a blicket/not a blicket?” for each block. After the ten blocks were demonstrated, children were asked to look at the “blicket” and “not blicket” boxes. Children were told that, “Most of the blocks we saw were not blickets. A few of them were, but almost all of the ones we tried were not blickets.” This was done to remind children about the base rate of an object being a blicket. This pretest and familiarization was identical to the procedure used by Sobel et al. (2004) in their “rare” condition.

Children were then given the test trials. In the first trial, the *AB-AC* trial, three blocks were taken out of the box, (A, B, and C). Two of them (A and B) were placed on the detector together, which activated. Then, one of those two blocks was placed on the detector with the block that had not been placed on the detector (A and C). The detector again activated. Children were then asked to categorize the block that was placed on the detector twice: “Which box does this one go in?” as well as the other two blocks: “What about these? Which box do these go in?” If children responded that they did not know, they were encouraged to

take a guess. Children were never allowed to place individual blocks on the detector. The spatial location of the blocks was counterbalanced.

Children were then given *a Baseline* trial. Two more blocks were brought out (X and Y). They were placed on the machine together, which activated. Children were asked to place these two blocks in the appropriate box.

Finally, a *Control* trial was done to ensure that children were on task. Two more blocks were brought out. Each was placed on the machine, one at a time. One made it go and one did not (randomly determined). Children were then asked to put the blocks into the appropriate box. If the child did not correctly categorize these blocks , they were not included in the analysis. Two children were excluded for this reason.

Results and Discussion

We will refer to the block placed on the machine twice as block A, the block placed on the machine with block A initially as block B, and the block placed on the machine with block A afterwards as block C. Table 4 shows the probability that children placed each block in the blicket box. Children differed in their overall treatment of blocks A, B, and C at the end of the AB-AC trial, Cochran's $Q(2, N = 16) = 6.22$, $p < .05$. Subsequent analysis showed that children did not differ in their treatment of blocks B and C, McNemar $\chi^2(1, N = 16) = 0.25$, *ns*., but they did differ in their treatment of block A and block C at the end of the trial, McNemar $\chi^2(1, N = 16) = 4.17$, $p < .05$ and differed in their treatment of block A vs. the combination of blocks B and C overall, Wilcoxon Signed Ranked Test, $z = -2.13$, $p < .05$, $r = 0.38$. Children did not differ in their treatment of block A and B at the end of the trial, McNemar $\chi^2(1, N = 16) = 1.13$, *ns.*

[Insert Table 4 Approximately Here]

Children categorized both blocks X and Y as blickets 87% of the time, and differed in their overall treatment of blocks B, C, X, and Y (i.e., all the blocks only shown to be effective once, always with another block), Cochran's Q (3, N = 16) = 10.13, $p < .05$. Specifically, they were more likely to categorize blocks X and Y as blickets than block C, both McNemar χ^2(1, N = 16)-values = 4.17, $p < .05$, and overall, they treated blocks X and Y together differently from blocks B and C together, Wilcoxon Signed Ranked Test, z = -2.43, $p < .05$, r = 0.43. However, they did not treat block X or Y significantly different from block B, both McNemar χ^2(1, N = 16)-values = 1.50, *ns*. Overall, however, these results qualitatively match the predictions of the model, and overall suggest that children integrated the prior probability information into their judgments.

These inferences are not easily reconcilable with the other alternatives we have considered. The DBR heuristic cannot explain why children were more likely to categorize object A as a blicket than objects B or C, and were less likely to categorize B and C as blickets at the end of the AB-AC trial than objects X and Y in the association trial. Children saw no unambiguous data about any of these objects that would support deductive reasoning about their efficacies and the base rate is approximately equal for all five objects. Similarly, while most associative learning models can account for the preference of object A over objects B and C, they fail to account for the fact that objects B and C are less likely to be categorized as blickets in the AB-AC trial than are objects X and Y in the Baseline trial. The strengths of association between these four objects and the detector's activation should be equal, because each object was observed to activate the detector once in the presence of another object.[7] In contrast, our Bayesian model correctly predicts all of the effects that we observed.

We do not conclude from these data that children necessarily have the same mechanism for causal inference as adults or are explicitly engaging in computations involving Bayes' rule. Rather, these data suggest that young children might have the ability to take into account information about the prior probability of particular kinds of causal relations when making judgments from ambiguous evidence.

Manipulating the Functional Form of Causal Relations

Experiments 1-3 suggest that adults and children are sensitive to the prior probability of existing causal relations, producing judgments that are quantitatively and qualitatively consistent with the predictions of our Bayesian model. But there are other ways that prior knowledge might influence new causal judgments. Our procedures require the learner to use another piece of more abstract knowledge beyond recognizing the base rate of objects with causal efficacy. We assume a deterministic activation law: a detector will only activate when an object with causal efficacy is placed on it, and it will always do so. In the next two experiments, we manipulate the participant's prior knowledge about the deterministic or probabilistic nature of the machine.

Assuming determinism allows adults and children to make strong inferences about causal relations from small amounts of data. Consider the predictions of our Bayesian model in a slightly different setting. Gopnik et al. (2001) presented preschoolers with similar blicket detector tasks. On their *one-cause* trials, children observed one object (A) that activated the detector by itself once. Then, children saw another object (B), placed on the detector, which did not activate. After B was removed, both A and B were placed on the detector together twice, and the detector activated both times. Having seen such a pattern of activation, 3- and 4-year-olds were confident that A was a blicket while B was not.

We can apply our Bayesian model to these trials in exactly the same way as in Experiment 1. The hypothesis space is identical and the assumptions outlined above provide a prior probability and a likelihood for each hypothesis. Observing the sequence of events in the *one-cause* trial produces the predictions that object A is definitely a blicket, while object B is definitely not, even with only four data points.

The key to drawing strong conclusions about the status of objects A and B is the deterministic nature of the activation law. Because the detector activates when object A is placed on it, Graph 0 and Graph 2 have a likelihood of 0. Likewise, because the detector does not activate when B is placed on it, Graph 3 has a likelihood of 0. The only causal structure with a non-zero likelihood is Graph 1, and consequently the posterior probability of that structure is 1, provided ρ is between 0 and 1. Applying Equation 4, we find that the probability that A is a blicket is 1, while the probability that B is a blicket is 0.

The deterministic activation law assumes that the machine will always activate in the presence of a blicket, and never activates in the absence of a blicket. But if the detector's mechanism is probabilistic instead of deterministic, we should make different assumptions.. One way we can instantiate this intuition by stating that the detector activates with probability ε when an object that is not a blicket is placed on it and activates with probability 1 - ε when a blicket is placed on it, where ε is a relatively small number. Under this theory, each object has an independent opportunity to activate the detector, meaning that if both objects are on the detector, there is a slightly higher probability that the detector activates (see the Appendix for details). This way of combining the causal strengths of the objects is known as a "noisy-OR" (Pearl, 1988), and is that assumed in the Power PC model (Cheng, 1997) as well as other

models based on Bayesian structure learning (Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001).

Suppose we assumed the faulty detector theory and then saw the one-cause or backwards blocking evidence. The probabilities of the evidence given that theory are shown in Table 5. Because the detector is more likely to activate in the presence of a blicket, object A is likely to be a blicket. However, the evidence against object B is no longer categorical, and some chance remains that object B is a blicket, modulated by the prior probability ρ, and the probabilistic parameter, ε.

[Insert Table 5 Approximately Here]

Allowing for a probabilistic mechanism in the detector raises a critical question: How does a learner know whether to adopt the deterministic or probabilistic assumption? This question can also be formulated as one of Bayesian inference, although the hypotheses involved are more abstract than specific causal structures. Now the relevant hypotheses are instances of causal theories about the nature of the activation law or causal mechanism that relates objects to the detector, which generate the specific causal models and priors that the learner considers to interpret the observed events (Griffiths & Tenenbaum, 2007; 2009; Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum, Griffiths, & Niyogi, 2007). For example, the perfect detector theory and the faulty detector theory both constitute hypotheses about the way that blicket detectors work, and both make predictions about the kind of events that one might observe involving blicket detectors. Bayes' rule can be used to select the hypothesis that provides the best account of the observed data. This computation is a kind of *hierarchical Bayesian inference*, and it is discussed in greater detail in the Appendix.

If we assume that only two theories are under consideration, a deterministic theory and a probabilistic theory, then the relevant Bayesian computations take on a particularly simple and intuitive form. The deterministic theory predicts that a block will always activate the machine, or else will never activate it. If the learner observes a single object activate the detector, and then fail to activate the detector, this theory is ruled out. So when people see this pattern of evidence they should favor a faulty detector theory, and make subsequent causal inferences based on this approach.

The next two experiments make use of this Bayesian framework to explore how adults (Experiment 4) and children (Experiment 5) use their observations to infer whether a causal mechanism is deterministic or probabilistic and then use this knowledge to make inferences about the causal efficacy of objects. We first present participants with data suggesting that a device operates via a deterministic mechanism (the perfect detector theory) or a probabilistic mechanism (the faulty detector theory). We then examine the inferences that participants make about new causal relations.

These experiments provide a further opportunity to differentiate the Bayesian account of causal structure learning from other accounts. The key prediction is that people will use their prior knowledge about mechanisms to draw different causal conclusions from exactly the same contingencies. Associative, causal strength, and deductive accounts (including the DBR heuristic) make inferences based purely on covariation between causes and effects, and thus cannot explain such a difference.

Experiment 4: Manipulating Functional Form in Adults

Method

Participants. The participants were 24 undergraduates recruited from a suburban-area university community.

Materials. A superlead detector similar to the one used in Experiments 1 and 2 was used here.

Procedure. Participants were introduced to the superlead detector and pencils in the same manner as in Experiments 1-2. They were then given a familiarization phase, in which they picked out two sets of three pencils to be scanned individually. Five out of the six pencils (randomly determined) would activate the detector.

Participants were randomly assigned to one of two conditions. In the *deterministic* condition, pencils either always or never activated the detector, and if a pencil activated the detector, it was labeled as containing super lead. Each pencil was scanned three times. Five of the six pencils activated the detector all three times, and were labeled as containing super lead; the other pencil failed to activate the detector all three times, and was labeled as not containing super lead. In the *probabilistic* condition, pencils activated the detector on 100%, 66%, 33%, or 0% of trials. If a pencil ever activated the detector, the pencil was labeled as containing super lead. One pencil activated the detector 100% of the time (3 of 3 times), two pencils activated the detector 66% of the time (2 of 3 times), and two other pencils activated the detector 33% of the time (1 of 3 times). These pencils were all labeled as containing super lead. The remaining sixth pencil failed to activate the detector all three times, and was labeled as not containing super lead. In the probabilistic condition, when a pencil activated the detector probabilistically, it always failed to activate it first, and then succeeded to do so. We did this to emphasize that objects that failed to activate the machine might still have efficacy.

After a pencil was labeled, it was placed to the side. Pencils were never re-used with the same participant.

The test phase began immediately after the familiarization trials, and was the same across the deterministic and probabilistic conditions. Participants were given three types of test trials (two of each type, six trials total). Each trial involved three pencils (A, B, and C), which were taken out of the cup by the participant and placed on the table in front of the detector. The spatial location of the three pencils was randomly determined, and the order of the type of trial was counterbalanced across participants.

In the *one-cause* trials, pencil A was placed on the detector, which activated. It was removed, and pencil B was placed on the detector, which did not activate. It was removed, and pencils A and B were placed on the detector together, which activated. This was demonstrated twice. Pencil C was then placed on the detector by itself. In one trial, it activated the detector, in the other, it did not.

In the *control-one* trials, pencil A was placed on the detector three times by itself activating it all three times. Pencil B was placed on the detector once, which failed to activate. Pencil C was placed on the detector twice: in one trial, it activated the machine, and on the other it failed to activate the machine.

In the *control-three* trials, pencil A was placed on the detector three times by itself, activating it all three times. Pencil B was placed on the detector three times, never activating the detector. Pencil C was placed on the detector once: in one trial, it activated the machine, and on the other it failed to activate the machine.

After each trial, participants were asked to rate the probability that each pencil contained super lead, using an 11-point scale from 0 to 10. Participants received the six trials

in one of six quasi-random orders. Following the Bayesian framework outlined above, there were two empirical questions. First, would people respond to pencil B in the *one-cause* trials differently across the deterministic and probabilistic conditions? In the deterministic condition, this pencil should clearly not contain super lead, in the probabilistic condition it might contain super lead.

Second, we were interested in whether there would be an interaction between ratings of pencil B between the control-one and control-three trials across the training conditions. For a learner with the perfect detector theory, the evidence that pencil B contains super lead is the same across the control-one and control-three trials. For a learner with the faulty detector theory, pencil B is more likely to contain super lead in the control-one trial than in the control-three trial. In the control-one trial, it is possible that pencil B has causal efficacy, and just failed to activate the detector on the one opportunity it had to do so, while in the control-three trial the detector would have to fail three times in a row. These predictions are shown in Figure 5: The probability that pencil B contains super lead is higher in the control-one trial than the control-three trial.

[Insert Figure 5 Approximately Here]

A third pencil, C, was always present (and either activated or failed to activate the detector) on every trial. This ensured that participants were not simply responding that only one pencil contained super lead on each trial. In both conditions, C should be treated as a super pencil if it activated the detector. In the deterministic condition, if C fails to activate the detector, it should not be a super pencil. In the probabilistic condition, it should be like the B object in the control-one trials.

Results and Discussion

Preliminary analyses revealed no differences between ratings for pencils A and B between the two repetitions of the one-cause, control-one, or control-three trials. As a result, the remainder of this analysis collapses the two repetitions together. The mean ratings of the probability that pencils A and B contained super lead for each trial and both conditions are shown in Figure 5. A 3 (Trial: *one-cause*, *control-one*, *control-three*) x 2 (Object: A vs. B) x 2 (Condition: *deterministic* vs. *probabilistic*) mixed Analysis of Variance was performed. Trial and Object were within-subject factors; Condition was a between-subject factor. A main effect of Trial was found, $F(2, 44) = 22.02$, $p < .001$, Partial $\eta^2 = 0.50$, as was a main effect of Object, $F(1, 22) = 1194.74$, $p < .001$, Partial $\eta^2 = 0.98$. A main effect of Condition was also found, $F(1, 22) = 11.58$, $p < .01$, Partial $\eta^2 = 0.35$. Several two-way interactions were significant, but these were subsumed by a significant three-way interaction among Trial, Object, and Condition, $F(2, 44) = 9.09$, $p < .001$, Partial $\eta^2 = 0.29$.

To minimize the risk of Type I error, instead of performing all simple effect analyses, we focused our subsequent analyses on the two empirical questions mentioned above. First, on the one-cause trials, did people respond differently about pencil B between the deterministic and probabilistic conditions? A *t*-test showed that there was a statistically significant difference in responses between the two conditions, with a mean rating of 0.58 in the *deterministic* condition and 3.75 in the *probabilistic* condition, $t(22) = 5.19$, $p < .001$, Cohen's $d = 2.12$. The higher level of ratings in the probabilistic condition is predicted by the Bayesian model.

Second, was there an interaction between the way participants categorized object B in the control-one and control-three trials across the two conditions? A 2 (Condition:

deterministic vs. *probabilistic*) x 2 (Trial: *control-one* vs. *control-three*) mixed Analysis of Variance was performed on responses to the test question for pencil B. Trial was a within-subject factor; condition was a between-subjects factor. There was a main effect of Trial, $F(1, 22) = 36.47$, $p < .001$, Partial $\eta^2 = 0.62$, and of Condition, $F(1, 22) = 26.46$, $p < .001$, Partial $\eta^2 = 0.55$, as well as a statistically significant interaction, $F(1, 22) = 13.13$, $p < .01$, Partial $\eta^2 = 0.37$. Participants in the *probabilistic* condition rated pencil B as more likely to contain super lead in the *control-one* trials than the *control-three* trials, with mean ratings of 4.25 and 0.92 respectively, while participants in the *deterministic* condition gave mean ratings of 1.00 and 0.17. This effect was also predicted by the Bayesian model.

We can use the procedure outlined in the Appendix, in which Bayesian inference is applied not just at the level of causal structures, but also at the level of causal theories, to make quantitative predictions about the results of this experiment. In order to do so, we need to fix values of ρ and ε. We set $\rho = 5/6$, because in the familiarization phase, five out of the six objects were super pencils. We treated ε as a free parameter, but set it to $\varepsilon = 0.1$, consistent with a prior that favors effective detectors, for which ε would be small. The resulting predictions are also shown in Figure 5. These settings of ρ and ε result in an extremely close quantitative fit, producing a linear correlation coefficient of $r = 0.996$ with the mean ratings. For comparison, the correlation between people's judgments and the causal power of A and B, computed using Equation 2, is $r = 0.956$. Since the Bayesian model has one free parameter, ε, while causal power has no free parameters, we examined the sensitivity of the Bayesian model to manipulation of this parameter. As shown in Figure 6, the Bayesian model produces a higher correlation than causal power provided the probability of the detector activating in the presence of a cause is high – a reasonable assumption about people's

expectations. More importantly, only the probabilistic model predicts the effect of familiarization with the probabilistic or deterministic detector on the rating of B.

[Insert Figure 6 approximately here]

The qualitative and quantitative correspondence between responses and the predictions of the Bayesian model support the claim that people are systematically and selectively applying prior knowledge about causal mechanisms when they learn about new causal relations. Learners seemed to be less certain about the status of pencil B when they assumed that the detector was probabilistic. The experiment also indicates that people can acquire beliefs about causal mechanisms from only a handful of examples (in our case, the behavior of the detector with six pencils). These results are at odds with both the associative learning and rational models of strength estimation, since they demonstrate that people reach different conclusions from exactly the same covariational information when they have different prior knowledge about the nature of causal relations.

Experiment 5: Manipulating Functional Form in Children

Our previous work with children suggested that their causal inferences are influenced by base-rate effects (see also Schulz, Bonawitz, & Griffiths, 2007). But there are no studies indicating whether children are also influenced by more abstract kinds of prior knowledge such as information about the deterministic or probabilistic nature of causes. Experiment 5 explores this question by replicating the procedure used in Experiment 4, using the blicket detector paradigm.

Method

Participants. The participants were 24 four-year-olds (13 girls, $M = 54.21$ months, range = 49-62 months) recruited from a local preschool and a list of hospital births. Five

additional participants were recruited, but were excluded from the study: four because of experimental error and one refused to participate.

Materials. The blicket detector used in Experiment 3 was used in this experiment. Twenty-four unique wooden blocks were divided into eight sets of three. In each set, no block was the same color or shape. Two small (approximately 2.5 cm in diameter) white porcelain knobs and two small metallic tee-joints (approximately 3cm in height) were also used.

Procedure. Children were administered the same warm-up as in Experiment 3 to ensure that children would interact with the experimenter, and accept that the experimenter would provide novel labels for objects. Children were then introduced to the blicket detector, and told that it was a "blicket machine" and that "blickets make the machine go." The remainder of the procedure closely followed that of Experiment 4. Half of the children were randomly assigned to the *deterministic* condition, and half of the children were assigned to the *probabilistic* condition, and given analogous familiarization with the detector to that provided to adults in Experiment 4. After each familiarization trial, children were asked to state whether each object was a "blicket." Corrective feedback was given if children answered incorrectly.

The test phase began immediately after the familiarization trials, and took the same form as that used in Experiment 4. Rather than providing ratings, each child gave a binary response for each object on each trial: They were asked whether each object was a blicket.

Results and Discussion

Only one child required corrective feedback on the familiarization trials, which suggested that children understood the basic structure of the procedure. In the trials where object C activated the detector, children claimed it was a blicket 93% of the time, and no

differently between the deterministic and probabilistic conditions. In the trials where object C failed to activate the detector, children claimed it was a blicket 7% of the time, and no differently between the conditions. Preliminary analyses also revealed no differences in the frequency of "yes" responses to the test question across the two one-cause, control-one, or control-three trials for either the A or B object, all McNemar $\chi^2(1) < 2.29$, all *ns*., so these data were combined. These data are shown in Table 6.

[Insert Table 6 Approximately Here]

A 3 (Trial: *one-cause*, *control-one*, *control-three*) x 2 (Object: *A* vs. *B*) x 2 (Condition: *deterministic* vs. *probabilistic*) mixed Analysis of Variance was performed. Trial and object were within-subject factors; condition was a between-subject factor. A main effect of Trial was found, $F(2, 44) = 15.41$, $p < .001$, Partial $\eta^2 = 0.41$, as was a main effect of Object, $F(1, 44) = 196.75$, $p < .001$, Partial $\eta^2 = 0.90$. A main effect of Condition was also found, $F(1, 22) = 16.59$, $p = .001$, Partial $\eta^2 = 0.43$. Several two-way interactions were significant, but these were subsumed by a significant three-way interaction between Trial, Object, and Condition, $F(2, 44) = 9.98$, $p < .001$, Partial $\eta^2 = 0.31$. This omnibus analysis revealed that differences in how the objects were categorized among the trials and between the conditions existed.

As in Experiment 4, our further analysis focused on the two empirical questions mentioned above. First, on the *one-cause* trials, did children respond differently about object B between the *deterministic* and *probabilistic* conditions? Responses to object B did differ between the two conditions, $t(22) = 7.60$, $p < .001$, Cohen's $d = 3.10$. Children were more likely to say object B was a blicket in the probabilistic condition (79% of the time) than in the deterministic condition (8% of the time). This suggests that the children recognized the difference between the two environments and reasoned accordingly.

Second, was there an interaction between children's' categorization judgments about object B in the *control-one* and *control-three* trials across the two conditions? A 2 (Condition: deterministic vs. probabilistic) x 2 (Trial: control-one vs. control-three) mixed Analysis of Variance was performed on responses to the test question for object B in these trials. Trial was a within-subjects factor; condition was a between-subjects factor. No main effect or significant interactions were found, unlike adults in Experiment 1, who showed a main effect of trial type as well as an interaction with condition. On average, children say that object B is a blicket only approximately 15% of the time on these trials across the two conditions.

The failure to find significant differences in the inferences about object B in the *control-one* and *control-three* conditions is inconsistent with our Bayesian model. However, this could reflect a genuine developmental difference: this task presented the most subtle inferences and the most taxing demands (with six test trials each involving three objects) of any blicket detector experiment we have run, and it would not be surprising if we have pushed four-year-olds beyond their abilities to approximate ideal Bayesian causal learners. However, the differences in results could also point to a disparity between the methods used with adults and children. Experiment 5 required children to make categorical responses (i.e., choosing whether each object is a blicket). In contrast, Experiment 4 allowed adults to use a more gradual rating scale. This methodological difference was unavoidable given our goal of obtaining quantitative judgments from adults that could provide a strong test of our model's subtle quantitative predictions. The coarser all-or-none responses required of children in Experiment 5 might have prevented these more subtle differences from emerging. Critically, though, the all-or-none response measure was sensitive enough to show that children recognized the difference in the one-cause trials between the two conditions – suggesting that

they were able to infer the general nature of the detector mechanism and to use that inference appropriately to guide some of their inferences about specific causal links.

General Discussion

Five experiments tested the hypothesis that human causal induction approximates rational Bayesian computations guided by appropriate forms of abstract prior knowledge. We tested two general predictions of this account. First, people are able to learn extremely quickly when they have appropriate prior knowledge, and the strength of the conclusions that they draw are determined by that knowledge – with high base rates and deterministic causes, they quickly become confident that causal relations exist – and by the extent to which the events they observe are ambiguous. Second, the nature of the conclusions that people reach is determined not just by covariation between cause and effect, but by how this covariation is interpreted in the light of prior knowledge. People's knowledge about the base rate that specific causal relations exist for a class of objects should influence how they interpret evidence about those objects. The graded effects of prior knowledge should be more pronounced with ambiguous evidence. Further, the nature of the mechanism underlying a causal relation – such as whether that mechanism is probabilistic or deterministic – should also influence causal judgments. These predictions set the Bayesian approach apart from other approaches to causal induction in which the evaluation of causal relations depends only on covariation between cause and effect. The results of our experiments suggest that people can make sophisticated and rational use of probabilistic reasoning in learning about causal systems.

In the remainder of the paper, we consider some of the implications of these results. First, we discuss whether these findings can be accounted for by other models. We then

briefly outline some of the possible developmental implications of our findings. Finally, we turn to the limitations of our analysis, and point out some possible directions for future work.

Can other models of causal learning account for these findings?

We will contrast the account offered by our Bayesian approach with several major competing traditions: models based on associative learning, rational estimators of causal strength parameters, and hybrids of deductive reasoning and simpler statistical heuristics.

Associative models. The present data present a challenge to accounts of causal learning based on associative mechanisms. While associative models might predict some of the trends we found, they do not predict the spectrum of learners' judgments across different levels of ambiguity in the evidence: from all-or-none inferences in the backwards blocking paradigm (Experiment 1) to more graded predictions in the AB-AC paradigm (particularly in Experiment 2). However, the greatest challenges for these models come from the effects of prior knowledge observed in our experiments. Since the contingencies between the prospective causes and the effect remained the same in all conditions of our experiments, a model of causal learning that is purely based on such contingencies cannot reproduce the effects we have observed.

We view this as an "in principle" argument – that associative models, generally construed, fail to capture our results because people's inferences vary when contingencies do not. It may be possible to modify these models in ways that make it possible to capture our results, at the cost of some parsimony. For instance, we see "belief revision" models (e.g., Catena, Maldonado, & Candido, 1998; Hogarth & Einhorn, 1992) as the most similar to the Bayesian framework we have proposed here. However, these models still fail to capture aspects of the present data. For instance, Catena et al. (1998) suggest that the belief in the

efficacy of a cause on a given trial is a function of whatever new evidence is observed on a given trial and the belief about that efficacy from previous trials, modulated by a learning parameter. In our experiments, the conclusions that adults and children reach are significantly changed by information about the overall probability that causal relations exist and whether or not those relations are deterministic, which are then applied to novel causes that have not been previously encountered. These inferences seem more sophisticated than those captured by existing "belief revision" models, although we could imagine modifications to these models that would fit the data. We place the burden of proof on advocates of such models to modify these accounts parsimoniously. Potential avenues to explore are providing an account of the effects of base rates in terms of shared features between the blocks used in establishing the base rates and the blocks used in subsequent inferences, and explaining why learning rate should be affected by information about whether causes are deterministic or probabilistic. We have not attempted to test these more complex models simply because the space of possible extensions to existing associative models is vast. We suspect that there are configurations of factors that would be able to predict the results of our experiments. Whether such a model is as parsimonious as the Bayesian account we have offered is an open question.

One final point worth exploring about associative models is that, in general, they were designed to make judgments from contingency information presented in trial-by-trial experiments, which is different from the inferences we asked of adults and children here. The assumption that the "activation law" makes is that there is a deterministic relation between objects containing super lead and activating the detector (or analogously being a blicket and activating the blicket machine). We do not directly test whether adults and/or children can make inferences about the strength of a causal relation based on differences in contingency

information (as in, e.g., Wasserman et al., 1993). We do, however, suggest that adults' inferences change as they are exposed to new data – for example, in Experiments 1 and 2, adults make different ratings about the likelihood of the objects having causal efficacy as they are exposed to each new data point presented in the test trial. Moreover, Danks, Griffiths, and Tenenbaum (2003) have demonstrated that a similar Bayesian model to the one we have presented here can account for differences in contingency information.

Power PC model. The power PC model (Cheng, 1997) and its extension to multiple causes (Glymour & Cheng, 1998; Cheng, 2000; Novick & Cheng, 2004) shares a deep commonality with the Bayesian model we have presented here: one underlying assumption of both models is that the interaction of multiple causes is essentially disjunctive. This means that an effect will occur if one or more of its potential causes are active. In fact, the power PC model can be thought of as a special case of inference over a causal graphical model, in which one assumes that the causal graphs to be learned have particular structure and parameterizations (see Glymour, 2001). In the Bayesian model, this assumption is embodied in the activation law: the blicket detector activates if one or more blickets are placed on top of it. In the power PC model, this assumption is embodied in the noisy-OR function that determines the probability of an effect conditioned on the presence or absence of potential causes (Glymour, 2001; Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001). Each cause is associated with a single strength parameter that determines how likely the effect is to occur if that cause – and only that cause – is active. The noisy-OR parameterization specifies how these strengths add up nonlinearly to determine the probability of the effect when multiple causes are active. When all causal strengths are equal to the maximal value of 1, the noisy-OR is equivalent to a logical OR function, and instantiates the assumption behind the

activation law. When the strengths are weaker than this maximal value, the presence of each additional cause increases the probability of the effect in proportion to its strength.

However, there are several key differences between our model and power PC. One is that our model attributes to learners a representation of the prior probability that a causal relation of a given type will hold. These priors may be calibrated based on the base rates of causal relations in the world, when learners have access to reliable base rate data, but may also reflect other sources of background knowledge. When calculating causal power, power PC relies on a calculation of the probability of the effect given the absence of the candidate cause, assuming that all of the potential causes are independent. This reflects a measure of the base rate of the effect occurring, but not in the same manner as the Bayesian model. On this interpretation, the causal power of any individual object is undefined before it has been placed on the detector. For all of the test objects in both of the present procedures, once they have been placed on the detector, their estimated causal powers should always be 1, because the detector always activates in the presence of the test objects. As such, power PC fails to predict a difference between the base rate conditions in Experiment 1, producing a correlation of $r = 0.325$, or the graded performance observed in Experiments 2 and 3, giving a correlation of $r = 0.768$ with the data from Experiment 2. Power PC gives better results for the manipulation of mechanisms used in Experiments 4 and 5, giving a correlation of $r = 0.956$ with the data form Experiment 4. However, it achieves this correlation without capturing the effect of familiarization with a probabilistic or deterministic detector on people's judgments.

A more charitable application of power PC to our experiments might assume that an object's default causal power – before it has been observed on the detector – is estimated to be the average causal power of the objects observed during training (which would be equivalent

to the base rate of objects activating the detector). This improves the performance of the power PC model in Experiment 1, giving a correlation of $r = 0.521$; performance on Experiments 2-5 is unaffected. Overall, however, these fits are still markedly worse than our Bayesian account.

A power PC account could be modified to include prior knowledge about causal power without losing its rational basis by using a maximum *a posteriori* probability estimate of the causal strength parameters (or the posterior mean), instead of a maximum likelihood estimate, or considering the posterior mean, as in Danks, Griffiths, and Tenenbaum (2003). Such a modification would allow the model to account for some but not all aspects of the data in these experiments. It would generate judgments that are modulated by the base rate of the effect, and that show some form of discounting, but that still do not fit the particular pattern of our results. The key difference between this model and our Bayesian account is in allowing objects to possess causal powers that vary continuously between 0 and 1. In our Bayesian account, an object is either a blicket (equivalent to possessing a causal power of 1), or is not a blicket (equivalent to possessing a causal power of 0). This assumption is critical to explaining how adults and children can draw inferences from such small samples: if intermediate values are allowed, many more observations would be needed to determine the true causal power of an object.

A recent variation on the Power PC theory instantiates some of the key assumptions that are embodied in our account of the knowledge guiding people's inferences about the blicket detector. Building on the work of Griffiths and Tenenbaum (2005), Lu et al. (2006, 2007, 2008) proposed that Bayesian inference over both causal structures and their strength can be used to evaluate causal relations. Instead of using a uniform prior over the strength of

causes, making each value of the causal power equally probable, as was done by Griffiths and Tenenbaum (2005), Lu et al. (2006, 2007, 2008) used a "necessary and sufficient" prior, favoring values of causal power closer to 0 and 1. They also added several additional terms to their prior, specific to modeling blicket detector tasks, which strongly favor objects as strong causes for the detector and weak strengths for background causes. These priors build into their model biases similar to the kinds of prior knowledge used in our model, and it may be possible to account for some of our results within their framework, since it shares the key elements of rational statistical inference and priors reflecting knowledge about the nature of causal relations. However, to account for the various manipulations we describe here, their priors would have to change accordingly, and it is not clear how natural the resulting account would be.

Heuristic accounts. Finally, we consider an alternative heuristic account, in which learners make deductive inferences given unambiguous information but default to using base rates when no other information is available (the DBR heuristic). The DBR heuristic can explain children's judgments in previous research and adults' performance in Experiment 1. Here, inferences about the B object are generated by incorporating base rate information with the deductive inference – given that two objects were chosen at random, and one is a super pencil, what is the likelihood that the other is a super pencil, given their base rate? There was some support for this approach in the 1/2 condition in Experiment 1: participants' inferences about the B object were slightly less than baseline (the Bayesian model predicts baseline performance). Participants might have reasoned that since the base rate of super pencils was 0.5, and one of these two objects clearly is a super pencil, the other must not be.

The procedure used in Experiments 2 and 3, however, was designed to test whether this alternative to the Bayesian account best described adults and children's inferences. In this procedure, no single object is ever unambiguously demonstrated as a cause, making such deductive inference impossible. The only way that the DBR heuristic could account for learners' inferences is if they chose to treat object A as a super pencil based on the data, and reason about object B and C based on base rate information. Adults' graded inferences in Experiment 2 are clearly inconsistent with this account, but on the surface, this might predict children's response pattern in Experiment 3. Children seem to treat object A as a blicket, and reduce their judgments about B and C appropriately (given that the base rate is low). However, this possibility is unlikely, given their treatment of objects X and Y – which are both often considered blickets. That is, the modal interpretation children make when shown two objects activate the detector together is that both objects are blickets – they do not deduce that only one is a blicket, and that the other must not be (a much more valid interpretation of these data, given the DBR heuristic).

The DBR heuristic also fails to account for the differences produced by providing evidence that the blicket detector employs a probabilistic or deterministic mechanism in Experiments 4 and 5. The success of the DBR heuristic in accounting for the results of Experiment 1 follows largely from the assumption that causal relations are deterministic, since this establishes a situation in which it is possible to obtain unambiguous evidence for the existence of a causal relation. As a consequence, the heuristic is not as useful when causal relations are probabilistic, and does not demonstrate the sensitivity to the difference between probabilistic and deterministic causal mechanisms exhibited by our participants in Experiments 4 and 5.

Implications for understanding the development of causal reasoning

Causal reasoning has been a topic of much interest in developmental psychology (e.g., Bullock et al., 1982; Carey, 2009; Leslie & Keeble, 1987; Shultz, 1982). While our goal was not to systematically explore the development of causal reasoning, our use of similar procedures with adults and children provides an opportunity to highlight some aspects of causal reasoning that seem similar in these two groups, and some that seem to change over time. Our formal framework also provides a starting point for a more detailed developmental exploration of how different aspects of children's knowledge about physical causal systems develop.

The results of Experiments 3 and 5 show that 4-year-olds behave in a way that is consistent with several of the predictions of our Bayesian model: they learn from small samples, are sensitive to the baserates with which causal relations exist, combine baserate information with observed evidence, and seem to reason differently about probabilistic and deterministic causal systems, as well as recognizing which kind of system they are dealing with in a specific context. However, Experiment 5 also revealed a way in which children seem to deviate from the predictions of our Bayesian model, not treating an object that fails to activate the detector once (in the *control-one* condition) differently from one that fails to activate the detector three times (in the *control-three* condition), and not modulating their interpretation of this evidence by whether the detector is deterministic or probabilistic.

The lack of a significant difference between the *control-one* and *control-three* conditions in Experiment 5 is a surprising finding. The statistical inference required to recognize the difference between these cases is one that appears in other aspects of learning: use of indirect negative evidence. When the detector is probabilistic, having more

opportunities to observe that a block fails to activate it provides stronger evidence against it being a cause, in the same way that the evidence that a construction does not belong to a language increases as a child obtains more linguistic input without hearing that construction. Use of indirect negative evidence plays an important role in many accounts of language acquisition (Pinker, 1979), and experiments in word learning with four-year-olds suggest that they can use such evidence in a way that is consistent with a Bayesian model (Xu & Tenenbaum, 2007a, 2007b). Further work thus needs to be done to determine whether the null effect observed in our experiment is the result of a lack of a domain-general ability to make such inferences, a result of limits on working memory, or simply a consequence of our experimental design.

Limitations and future directions

Our goal in this paper was to present a detailed test of the predictions of a specific Bayesian model of causal learning, allowing us to explore how prior knowledge influences the conclusions that adults and children reach about causal structure. By adopting such a specific focus, our analysis has several limitations, which provide opportunities for further research.

First, we focused on a specific kind of causal system – a physical system that acts as a "detector" of a causal property. This approach has both strengths and weaknesses. The strength is that we were able to examine aspects of prior knowledge specific to this kind of system, rather than using a more generic setting in which prior knowledge might be more diffuse. The weakness is that our analysis is limited to this case, with the expectation that other models and other forms of prior knowledge will be necessary in other cases. We expect future work to extend the scope of this Bayesian approach to causal induction beyond the

blicket detector paradigm. While we have used this paradigm as the basis for an in-depth exploration of the predictions of this account, the basic principle of using prior knowledge to guide statistical inferences about causal relations is one that can be applied more generally, and should be tested more broadly. Some recent work has already pushed the boundaries of this approach. Sobel and Munro (2009) have shown that children's inferences about psychological states nicely fits with this Bayesian model, and Schulz et al. (2007) illustrated how a similar approach can be used to investigate the effects of intuitive theories on causal learning across the biological and psychological domains. Kushnir and Gopnik (2007) similarly found that children were more willing to override spatial contiguity assumptions in a probabilistic than a deterministic context, and also suggested a Bayesian process of integrating prior knowledge and current evidence. Griffiths and Tenenbaum (2007, 2009) provide a formal framework intended to make it possible to extend the kind of analysis we present here to other, richer, causal systems.

One consideration that arises in extending this approach to other causal systems is the question of how far we might expect the correspondence between Bayesian models and human behavior to persist as hypotheses about causal structure become more complex. A limitation of the work we present here, and the related work summarized in the previous paragraph, is that even our experiments with adults involve reasoning about a small number of causes with no unknown factors. It is an open question whether people will continue to behave in a way that is consistent with Bayesian inference in the face of much more complex data, since such rational models assume perfect memory for the data and a large hypothesis space. Recent work examining adult causal learning in more complex settings suggests that

people might take a more "piecemeal" approach to assembling causal structures (Fernbach & Sloman, 2009).

A second limitation of our analysis is that we have focused on only one level at which computational models of causal reasoning might be defined – the computational level. As a consequence, our Bayesian model makes clear our assumptions about the prior knowledge that informs people's inferences, but it does not make precise predictions about cognitive processing steps that might implement or approximate these computations. Connecting the computational and algorithmic levels is a general challenge for Bayesian models of cognition (Bonawitz & Griffiths, 2010; Sanborn, Griffiths, & Navarro, 2010), but seems particularly relevant in the case of causal learning. Further elaboration of the model in this direction could provide some insight into the developmental trajectory of causal learning, or the more complex conditions under which adults sometimes fail to make successful causal inferences in the real world.

Finally, a concern that naturally arises when new computational models are introduced is whether those models are falsifiable. In particular, the Bayesian approach has a great deal of flexibility in the assumptions that are made about priors and hypothesis spaces, seeming to create the opportunity to fit a wide range of results. In considering this question, it is worth making a distinction between the Bayesian approach to modeling, and a specific Bayesian model. The Bayesian approach, like other broad computational frameworks such as production systems (Anderson, 1993) or connectionism (Rumelhart, McClelland, & the PDP Research Group, 1986), is not something we should expect to directly test empirically. The criteria for evaluating such frameworks is whether they lead to useful insights about cognition. However, specific models we should expect to be able to falsify. A model makes

commitments about hypothesis spaces and priors that lead to direct predictions, and behavior inconsistent with these predictions provides evidence against that model. The model we present here – with a set of principles that determine the hypothesis space and prior – is certainly falsifiable. In all of our experiments, there are quite reasonable alternative patterns of data that would be inconsistent with our model that could have emerged but did not (e.g., not attending to baserates in Experiment 1, judging all objects to be effective in Experiments 2-3, treating object B as ineffective across both conditions in Experiments 4-5). Further work will be needed to determine the explanatory scope of this model, but we do not anticipate that there will be any difficulty in falsifying it in contexts where it is not appropriate.

Conclusion

Combining prior knowledge with observed data is a critical part of causal learning, and the key to being able to make rapid causal inferences. Bayes' theorem provides the basis for a rational analysis of such inferences, and a framework for characterizing the prior knowledge that makes them possible. We have presented five experiments testing a Bayesian model of causal learning in children and adults. This model interprets observed data by applying rational statistical inference mechanisms to a hypothesis space of candidate causal structures, a space based on knowledge about the kinds of mechanisms relating causes to effects and the prior probabilities of encountering causal relations of various types. The model makes precise predictions about a range of effects, including use of baserate information, maintenance of graded degrees of belief, and the effects of exposure to evidence that the mechanism underlying a causal relation is probabilistic or deterministic. These predictions were confirmed both quantitatively in adults and qualitatively in child learners. We view these

results as a first step towards a more complete account of the prior knowledge that informs human causal reasoning across the wide range of domains in which it takes place.

References

Ahn, W., & Kalish, C.W. (2000). The role of mechanism beliefs in causal reasoning. In F.C. Keil and R.A. Wilson (eds.), *Explanation and Cognition* (pp. 199-226). Cambridge, MA: MIT Press.

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

Anderson, J. R. (1993). *Rules of the mind.* Hillsdale, NJ: Erlbaum.

Bonawitz, E. B., & Griffiths, T. L. (2010). Deconfounding hypothesis generation and evaluation in Bayesian models. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.*

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209-254). New York: Academic Press.

Carey, S. (2009). *The origin of concepts*. Oxford: Oxford University Press.

Cartwright, N. (1989). *Nature's capacities and their measurement*. New York: Oxford/ Clarendon Press.

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *17*, 837-854.

Chater, N., & Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*, 287-291.

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

Cheng, P.W. (2000). Causal reasoning. In R. Wilson & F. Keil (Eds.), The MIT Encyclopedia of Cognitive Sciences (pp. 106-108). Cambridge, MA: Bradford, MIT Press.

Christensen- Szalanski, J. J. & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 928-935.

Cole, R. P., Barnet, R. C., & Miller, R. R. (1995). Effect of relative stimulus validity: Learning or performance deficit*? Journal of Experimental Psychology: Animal Behavior Processes*, *21*, 293-303.

Cooper, G. F. (1999) An overview of the representation and discovery of causal relationships using Bayesian networks. In: C. Glymour & G. F. Cooper (Eds.), *Computation, Causation, and Discovery.* Cambridge, MA: Bradford, MIT Press.

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

Danks, D, Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (p. 35-42). Cambridge, MA: MIT Press.

Dickinson, A. (2001). Causal learning: Association versus computation. *Current Directions in Psychological Science, 10*, 127-132.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology*, *49B*, 60-80.

Dickinson, A., & Shanks, D. R. (1995). Instrumental action and causal representation. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp.5-25). Oxford: Oxford University Press.

Fernbach, P. M. & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 678-693.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In D. H. Fisher (ed.), *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997)* (pp. 125-133). San Francisco, CA: Morgan Kaufmann.

Friedman, N., & Koller, D. (2003). Being Bayesian about Bayesian network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, *50*, 95-125.

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

Glymour, C. (2003). Learning, prediction, and causal Bayes nets. *Trends in Cognitive Science, 7,* 43-48.

Glymour, C. & Cheng, P.W. (1998). *Causal mechanism and probability: A normative approach*. In Y. Oaksford and V. Chater (Eds.), Rational Models of Cognition (pp. 295-313). Oxford, England: Oxford University Press.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25*, 565-610.

Gopnik, A. & Glymour, C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation. In P. Carruthers, S. Stich, & M. Siegal

(Eds.), *The cognitive basis of science* (pp. 117-132) Cambridge: Cambridge University Press.

Gopnik, A. Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T, & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 1-30.

Gopnik, A., & Schulz, L. E. (2007). *Causal learning: Psychology, philosophy, computation*. New York: Oxford University Press.

Gopnik, A., Sobel, D. M., Schulz, L. & 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*, 620–629.

Green, D.M., & Swets J.A. (1966) *Signal detection theory and psychophysics.* New York: Wiley.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 285-386.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17*, 767-773.

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

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

Goldvarg, E. & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25,* 565-610.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. Jordan (Ed.), *Learning in Graphical Models.* Cambridge, MA: MIT Press.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational Models of Cognition* (pp. 248-274). Oxford University Press.

Holyoak, K. J., Koh, K., & Nisbett, R. E. (1989). A theory of conditioning: Inductive learning within rule-based default hierarchies. *Psychological Review,* 96, 315-340.

Hume, D. (1978). *A treatise of human nature.* Oxford: Oxford University Press. (Original work published 1739).

Jordan, M. I. (1998). *Learning in Graphical Models*. Cambridge, MA: MIT Press.

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition*, *50*(1-3), 189-209.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237-251.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107-128.

Kemp, C. & Tenenbaum, J. B. (2003). Theory-based induction. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society.*

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review, 7*, 636-645.

Krynski, T. R. & Tenenbaum, J. B. (2003). The role of causal models in reasoning under uncertainty. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society.*

Krynski, T. R. and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 136*, 430-450.

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

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology, 43*, 186-196.

Larkin, M. J. W., Aitken, M. R. F., & Dickinson, A. (1998). Retrospective revaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 1331-1352.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition, 25*, 265-288.

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

Lovibond, P. F., Been, S., Mitchell, C. J., Bouton, M. E. & Frohardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory & Cognition, 31*, 133-142.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2006). Modeling causal learning using Bayesian generic priors on generative and preventive powers. In R. Sun & N. Miyake (Eds.), *Proceedings of the Twenty-eighth Annual Conference of the Cognitive Science Society* (pp. 519-524). Mahwah, NJ: Erlbaum.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2007). Bayesian models of judgments of causal strength: A comparison. In D. S. McNammara & G. Trafton

(Eds.), *Proceedings of the Twenty-ninth Annual Conference of the Cognitive Science Society* (pp. 1241-1246). Mahwah, NJ: Erlbaum.

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

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford: Oxford University Press.

Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. *Journal of Personality and Social Psychology, 38,* 231-241.

McCormack, T., Butterfill, S., Hoerl, C., & Burns, P. (2009). Cue competition effects and young children's causal and counterfactual inferences. *Developmental Psychology*, *45,* 1563-1575.

Miller, R. R., & Matute, H. (1996). Biological significance in forward and backward blocking: Resolution of a discrepancy between animal conditioning and human causal judgment. *Journal of Experimental Psychology: General*, *125*, 370-386.

Mitchell, C. J., Killedar A., & Lovibond, P. F. (2005). Inference-based retrospective revaluation in human causal judgments requires knowledge of within-compound relationships. *Journal of Experimental Psychology: Animal Behavior Processes, 31*, 418-424.

Morris, M. W. & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, *102*(2), 331-355.

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

Oaksford, M. & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608-631.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychology Review, 94*, 61-73.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo; CA: Morgan Kaufman.

Pearl, J. (2000). *Causality*. New York: Oxford University Press.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. New York: W. W. Norton.

Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*, 217-282.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current theory and research* (pp. 64-99). New York: Appleton-Century-Crofts.

Rumelhart, D. E., McClelland, J. L., & the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world.* Princeton: Princeton University Press.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117*, 1144-1167

Schlottman, A. (2000). Children's judgments of gambles: A disordinal violation of utility. *Journal of Behavioral Decision Making, 13*, 77-89.

Schulz, L.E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared make your tummy ache? Naive theories, ambiguous evidence and preschoolers' causal inferences. *Developmental Psychology*, *43*, 1124-1139.

Seiver, E., Gopnik, A., Lucas, C., & Goodman, N. D. (2007). *Causal inference and the origins of social cognition – preschool children use covariation to make trait attributions.* Poster presented at the Meeting of the Society for Research in Child Development, Boston, 2007.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology, 40*, 162-176.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the society for research in child development, 47(1)*, 1-51.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgment. *Quarterly Journal of Experimental Psychology, 37B*, 1-21.

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 48*, 257-279.

Sloman, S., & Lagnado, D. A. (2005). Do people 'do'. *Cognitive Science,* 29, 5-39.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology, 42*, 1103-1115.

Sobel, D. M., & Kirkham, N. Z. (2007). Bayes nets and Babies: Infants' developing representations of causal knowledge. *Developmental Science, 10,* 298-306.

Sobel, D. M., & Munro, S. A. (2009). Domain generality and specificity in children's causal inferences about ambiguous data. *Developmental Psychology*, 45, 511-524.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28,* 303-333.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search (Springer Lecture Notes in Statistics, 2nd edition, revised)*. Cambridge, MA: MIT Press.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. Solla, T. Leen, and K. R. Mueller (eds.), *Advances in Neural Information Processing Systems 12*. Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001a). Structure learning in human causal induction. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (p. 59-65). Cambridge, MA: MIT Press.

Tenenbaum, J. B. & Griffiths, T. L. (2001b). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*(4), 629-641.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (p. 35-42). Cambridge, MA: MIT Press.

Tenenbaum, J. B. and Xu, F. (2000). Word learning as Bayesian inference. *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*.

Tversky, A. & Kahneman, D. (1980). Causal Schemata in Judgments Under Uncertainty. In M. Fishbein (Ed.), Progress in Social Psychology, vol. 1, (pp. 49-72). Hillsdale, NJ: Erlbaum.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25*, 127-151.

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), The psychology of learning and motivation, Vol. 34: Causal learning (pp. 47-88). San Diego: Academic Press.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82*, 27-58.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing Versus Doing: Two Modes of Accessing Causal Knowledge. *Journal of Experimental Psychology: Learning Memory and Cognition, 31,* 216-227.

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

Wasserman, E. A., & Berglan, L. R (1998). Backward blocking and recovery from overshadowing in human causal judgment: The role of within-compound associations. *Quarterly Journal of Experimental Psychology: Comparative & Physiological Psychology, 51*, 121-138.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*, 82-111.

Xu, F. & Tenenbaum, J. B. (2007a). Sensitivity to sampling Bayesian word learning. *Developmental Science, 10*, 288-297.

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

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Science*, *10*, 301-308.

Footnotes

1. Developmental differences were also found: 3-year-olds did not respond in this manner – they did not seem sensitive to the rare-common manipulation. Sobel et al. (2004) discuss a number of potential reasons for this developmental difference, and preliminary evidence (Sobel & Munro, 2009) suggests that 3-year-olds' inferences are also consistent with a Bayesian account constrained by different pieces of prior knowledge.
2. We refer the reader to Glymour (2003) or Gopnik et al. (2004) for relatively accessible descriptions of causal graphical models and the Causal Markov condition. Glymour (2001) and Pearl (2000) provide a more in-depth description of these topics.
3. In fact, the hypothesis space for any blicket detector experiment has to involve arbitrary numbers of objects, since new objects with unknown causal powers could be introduced at any point. This does not present a problem for our model: since we assume that the probability that each object is a blicket is independent of all other objects, unobserved objects do not influence conclusions about observed objects. Consequently, we can work with a hypothesis space in which we represent only the causal relations among the observed objects. More formally, each hypothesis can be considered to correspond to an equivalence class of all hypotheses that include unobserved objects in which these are the causal relations involving observed objects.
4. The approach we take in this section is formally equivalent to another Bayesian analysis, in which we have a random variable indicating whether each object is a blicket and perform probabilistic inference to identify the categories of these objects. Since belonging to the category of blickets corresponds exactly with having the ability

to cause the detector to activate, the hypothesis space over causal structures is equivalent to the hypothesis space over category memberships. A Bayesian analysis more along these lines is presented by Griffiths and Tenenbaum (2007).

5. Where we report causal power in the experiments, it was computed using the contingency table for each potential cause with the effect separately, aggregated over all trials. When there were no observations of the effect in the absence of the cause this probability was taken to be zero, and undefined values of causal power were also set to zero. Used in this way, the model does not naturally predict blocking (either forwards or backwards) in the strong sense that the Bayesian model does, since it still estimates a positive causal power for B. However, the causal power of B is less than that of A since the trials on which the effect occurred in the presence of A alone increase the probability of the effect in the absence of B. We include this simple model primarily to illustrate that predicting the results of the experiments is non-trivial, and to provide a quantitative comparison for the reported correlations.
6. Interestingly, the average rating of object C at this point was 1.90, which tended to be lower than the baseline response: $t(19) = 1.91$, $p < .10$. This tendency could reflect a version of the "gambler's fallacy": if super pencils are rare and the probability that two of the three objects are super pencils increases, then the third object may be judged less likely to be a super pencil than its baseline probability dictates as a kind of correction. Alternatively, this decrease could simply reflect a pragmatic inference that participants make: given the task setup, most likely at least one object is a super pencil and at least one object is not. Given the increased evidence for the other two objects being super pencils, the third object is most likely to be the non-super pencil. We will

not pursue this effect here, other than to note that it is not nearly as strong as the other effects we report, and that its origin is an interesting question for future work.

7. This is particularly true for object B, which should have exactly the same associative strength as objects X and Y. On many associative models, object C should have slightly lower associative strength than object B, since object C was paired with object A, which already accrued associative strength when it was first placed on the detector with object B.

Author Note

The first three authors contributed equally to this work and their order is alphabetical. This research was made possible by grants from NIH (F31-MH12047 to DMS) and NSF (DLS-0132487 to AG and DLS-0518161 to DMS), funding from Mitsubishi Electric Research Labs and the Paul E. Newton chair to JBT, grants from the Air Force Office of Sponsored Research to JBT and TLG, and a grant from the J. S. McDonnell Foundation Causal Learning Collaborative. We would like to thank Anu Asnaani, Brian Christian, Alana Firl, and Ellen Hamilton for assistance in data collection, and Clark Glymour, Tamar Kushnir, and Laura Schulz for helpful discussion.

Figure Captions

Figure 1. Causal graphical models indicating the possible causal relations for events involving two objects and one detector. *A* and *B* indicate the presence of objects A and B on the detector, and *E* indicates the activation of the detector.

Figure 2. Predictions of the Bayesian model for each condition (top graphs) and mean ratings of participants in each condition (bottom graphs) in Experiment 1.

Figure 3. Causal graphical models indicating the possible causal relations for events involving three objects and one detector. *A*, *B*, and *C* indicate the presence of objects A, B, and C on the detector, and *E* indicates the activation of the detector.

Figure 4. Predictions of the Bayesian model for each condition (left) and mean ratings of adult participants in each condition (right) in Experiment 2.

Figure 5. Adults' ratings of the likelihood that objects A and B are super pencils across the probabilistic and deterministic conditions in Experiment 4 and the corresponding predictions of the Bayesian model.

Figure 6. Correlation of the Bayesian model with human responses as a function of the strength of the probabilistic cause, corresponding to 1- ε. For a range of values of ε consistent with relatively strong causal relationships, this correlation is greater than the correlation produced by causal power, shown with the dotted line.

Appendix

The hypotheses shown in Figure 1 represent parameterized causal graphical models relating three variables: the presence of object A on the detector (A), the presence of object B on the detector (B), and the activation of the detector (D). The deterministic activation law specifies the parameterization of these causal graphical models, with the probability that the detector activates being given by

$$p(e+ \mid A, B) = 1 - (1 - A)^{(A \rightarrow E)} (1 - B)^{(B \rightarrow E)} \qquad \text{(A1)}$$

where we assume A and B take on the values 1 when true and 0 when false, and $A \rightarrow E$ and $B \rightarrow E$ denote the presence of a link from A and B to E respectively. The resulting probability is 1 if one of the causes of E is present and 0 otherwise.

The prior probability of each of the four hypotheses follows from the principle of object independence, and is given in Table 2. As stated in Equation 4, the probability that object A is a blicket is equal to the probability that the $A \rightarrow E$ link exists, and likewise for object B. These probabilities are equal to the sum of the probabilities of all hypotheses in which that link exists. It is straightforward to check that the prior probability that A and B are blickets is equal to ρ, in accord with the principle of object independence, as recorded in Table 2.

The backwards blocking paradigm has two stages. The first stage consists of some number of AB events, in which both A and B are placed on the detector and the detector activates. We can encode this the event $e+|a+,b+$, indicating that E was present when A and B were set to be present. We need not model the probabilities of A and B, since these variables were set to their values by an external intervention (Pearl, 2000). Taking such an event as our data d, we need to compute the probability of d under each hypothesis h in order

to apply Bayes' rule (Equation 3). It follows from the activation law that $p(d \mid h) = 1$ if the graph corresponding to h contains at least one of $A \rightarrow E$ and $B \rightarrow E$, and 0 otherwise. Consequently, applying Bayes' rule simply involves identifying which hypotheses are consistent with the data, and computing posterior probabilities by re-normalizing the prior probabilities of each of those hypotheses by the sum of the prior probabilities of the members of that set. Three hypotheses are consistent with the event $e+|a+,b+$: Graph 1, Graph 2, and Graph 3. The sum of the prior probabilities of these hypotheses is $\rho(1-\rho) + \rho(1-\rho) + \rho^2 = \rho(2-\rho)$. Dividing the prior probability of each hypothesis by this quantity gives the posterior probabilities shown in Table 2. The posterior probability of A and B being blickets can be computed by summing over those graphs in which the appropriate causal relations exists. These posterior probabilities remain the same regardless of the number of instances of $e+|a+,b+$ in d, since the set of hypotheses consistent with such events remains unchanged.

In the second stage of the backwards blocking paradigm, participants see an A event, in which A is placed on the detector alone and the detector activates. In terms of our variables, this is the event $e+|a+,b-$. Following the activation law, such an event has probability 1 if a causal structure contains $A \rightarrow E$, and probability 0 otherwise. If d consists of the two events $e+|a+,b+$ and $e+|a+,b-$, then $p(d \mid h)$ is the product of the probability of each of these two events under h, which is also 1 if a causal structure contains $A \rightarrow E$, and 0 otherwise. Only two hypotheses contain $A \rightarrow E$ – Graph 1 and Graph 3 – so the posterior probability is the prior probability re-normalized over this set. The sum of the prior probability of these hypotheses is $\rho(1-\rho) + \rho^2 = \rho$, and dividing the prior probabilities by this amount yields the posterior probabilities shown in Table 2.

While we have focused on the case of two objects, similar principles can be used to compute posterior probabilities for the hypotheses for three objects shown in Figure 3: the activation law rules out all hypotheses inconsistent with the observed data, and the posterior probabilities are a re-normalized version of the prior probabilities. This procedure was used to compute the posterior probabilities shown in Table 3.

Under a probabilistic activation law, in which non-blickets activate the detector with probability ε and blickets activate the detector with probability 1- ε, the probabilità of the detector activating is given by

$$p(e+ \mid A,B) = 1 - \varepsilon^{N_{blicket}} (1-\varepsilon)^{N_{total} - N_{blicket}} \qquad \text{(A2)}$$

where N_{total} is the total number of objects on the detector and $N_{blicket}$ is the number of these that are blickets. It is straightforward to check that this yields the probabilities given in Table 5. The predictions of the model were obtained by applying Bayes' rule, calculating the probabilities of the events in the *one cause*, *one control*, and *three control* conditions under the four hypothetical causal structures, and combining them with the same prior as used for the deterministic detector.

We can also use Bayesian inference to choose between qualitatively different causal theories. If we use T_P to denote the "perfect detector" theory and T_F to denote the "faulty detector" theory, then we can apply Bayes' rule, with

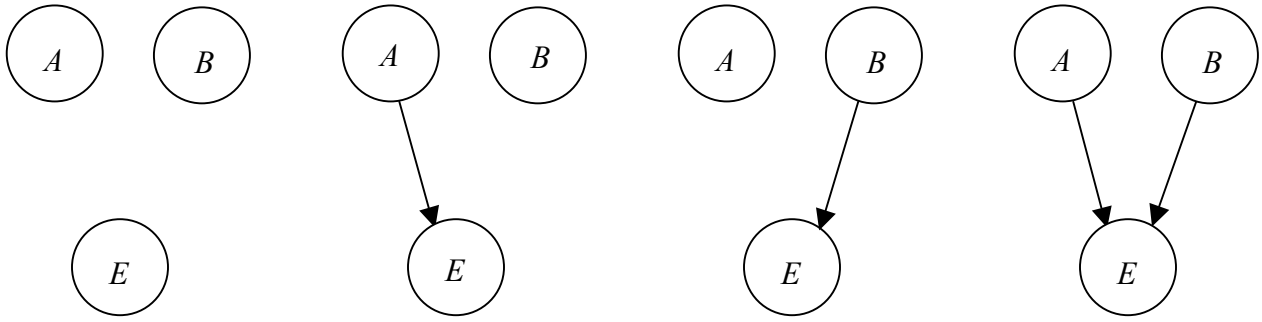
$$P(T_P \mid d) = \frac{P(d \mid T_P)P(T_P)}{P(d \mid T_P)P(T_P) + P(d \mid T_F)P(T_F)} \qquad \text{(A3)}$$

where the likelihood, $P(d \mid T)$, indicates the probability of some sequence of events under theory T, and the prior, $P(T)$, represents the *a priori* plausibility of that theory. The likelihood is computed by summing over all possible causal structures,

$$P(d \mid T) = \sum_{h \in H} P(d \mid h, T) P(h \mid T) \qquad \text{(A4)}$$

where $P(d \mid h, T)$ and $P(h \mid T)$ are just the likelihood and prior used in computing the posterior distribution over causal structures assuming a particular theory. For the "perfect detector" and "faulty detector" theory, the outcome of this inference depends on whether the observed data obey the activation law under any of the candidate causal structures. If every trial respects the deterministic activation law for some candidate causal structure, the "perfect detector" theory will ultimately win out, as the "faulty detector" theory predicts failures on at least some trials. However, if there is ever an observation which is inconsistent with the deterministic activation law applied to any candidate causal structure – such as observing a single object that both activates and fails to activate the detector – the "perfect detector" theory will immediately be assigned likelihood 0 (and hence posterior probability 0), and the "faulty detector" theory will immediately obtain a posterior probability of 1.

Figure 1



Graph 0 Graph 1 Graph 2 Graph 3

Figure 2

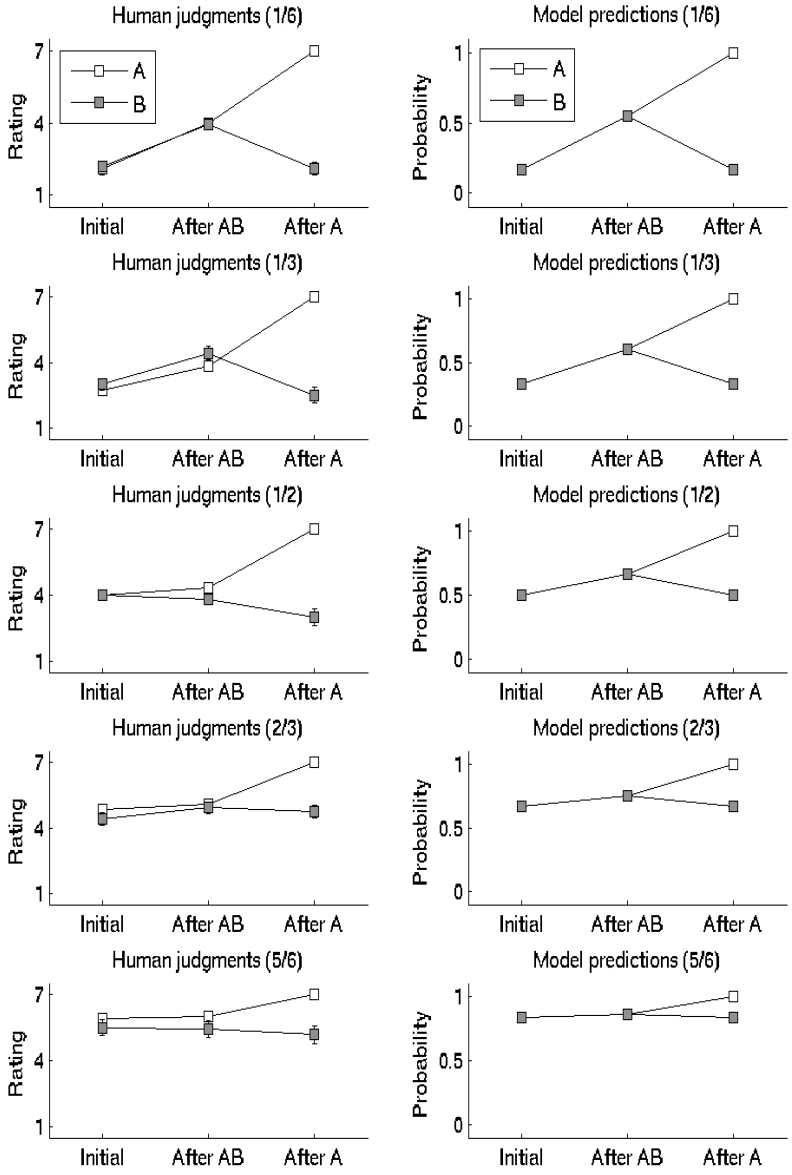
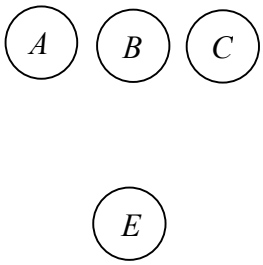
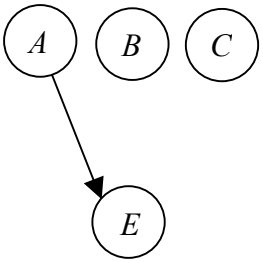

Human judgments (1/6)
Model predictions (1/6)
Human judgments (1/3)
Model predictions (1/3)
Human judgments (1/2)
Model predictions (1/2)
Human judgments (2/3)
Model predictions (2/3)
Human judgments (5/6)
Model predictions (5/6)
Rating
Probability
A
B
Initial
After AB
After A

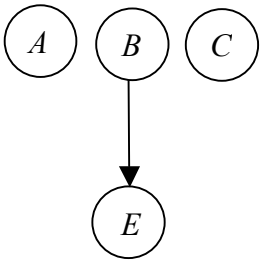
Figure 3



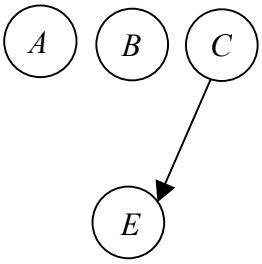
Graph 0



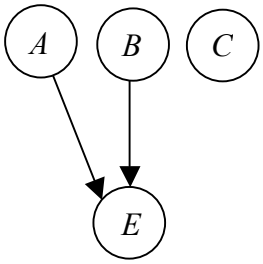
Graph 1



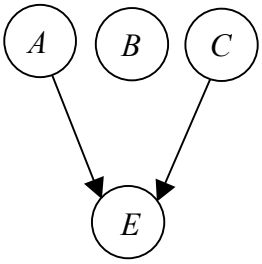
Graph 2



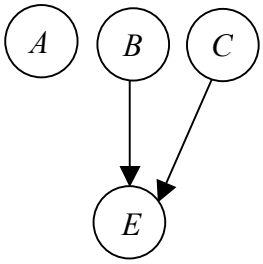
Graph 3



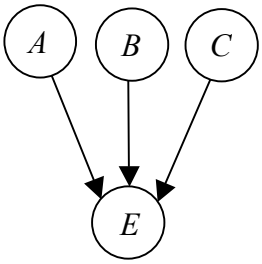
Graph 4



Graph 5



Graph 6



Graph 7

Figure 4

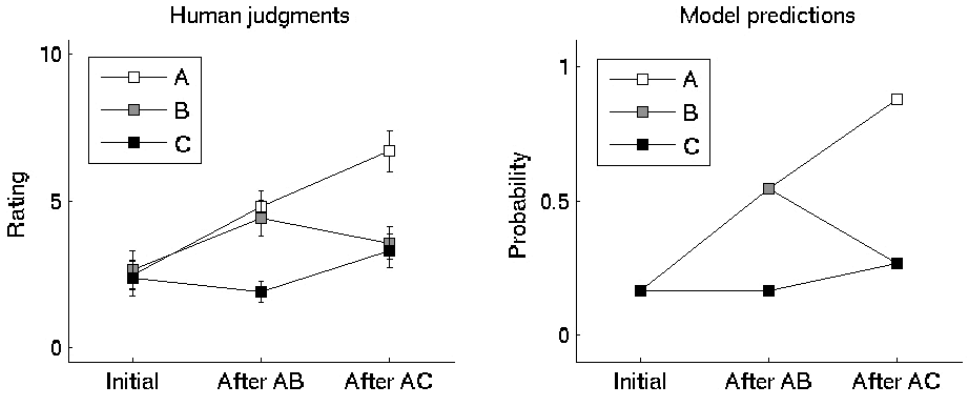

Human judgments
Rating
10
5
0
A
B
C
Initial
After AB
After AC
Model predictions
Probability
1
0.5
0
A
B
C
Initial
After AB
After AC

Figure 5

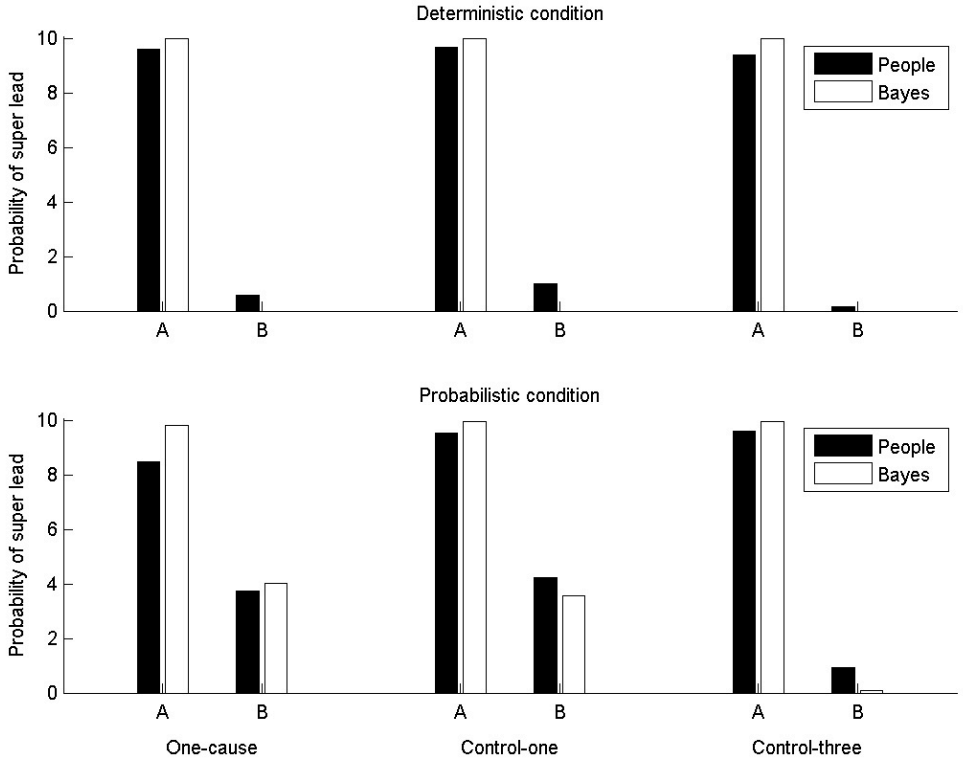

Deterministic condition
Probability of super lead
10
8
6
4
2
0
People
Bayes
A
B
A
B
A
B
Probabilistic condition
Probability of super lead
10
8
6
4
2
0
People
Bayes
A
B
A
B
A
B
One-cause
Control-one
Control-three

Figure 6

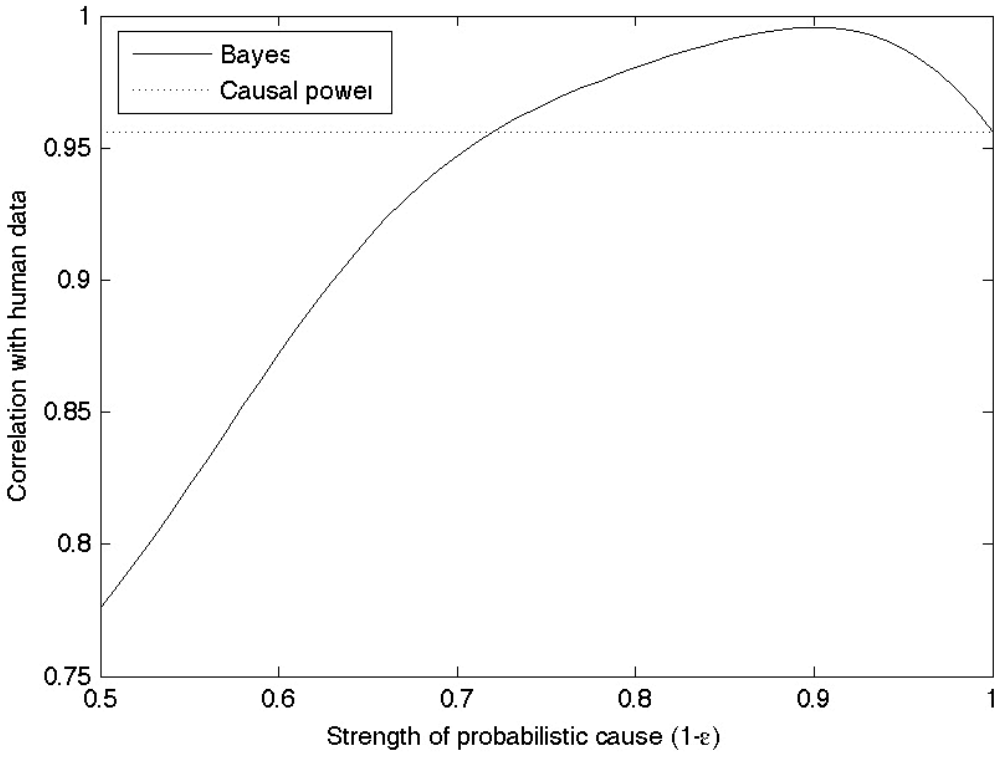

Bayes
Causal power
Correlation with human data
Strength of probabilistic cause (1-ε)
1
0.95
0.9
0.85
0.8
0.75
0.5
0.6
0.7
0.8
0.9
1

Table 1

Probability of Different Events for each Causal Structure with Deterministic Activation Law

	Event		
	e+|*a*+,*b*+	*e*+|*a*+,*b*-	*e*-|*a*-,*b*+
Causal Structures			
Graph 0	0	0	1
Graph 1	1	1	0
Graph 2	1	0	1
Graph 3	1	1	0

Note. The notation *e*|*a,b* indicates the state of the effect (activation of the detector) given the state of the causes (objects A and B being on the detector), with + indicating presence and – indicating absence.

Table 2

Posterior Probabilities for Bayesian Model of Backwards Blocking (Experiment 1)

	Prior Probability	After AB Event	After A Event
Causal Structures			
Graph 0	$(1-\rho)^2$	0	0
Graph 1	$\rho(1-\rho)$	$(1-\rho)/(2-\rho)$	$1-\rho$
Graph 2	$\rho(1-\rho)$	$(1-\rho)/(2-\rho)$	0
Graph 3	ρ^2	$\rho/(2-\rho)$	ρ
Probability of Being a Blicket / Containing Super Lead			
Object A	ρ	$1/(2-\rho)$	1
Object B	ρ	$1/(2-\rho)$	ρ

Note. The probability of an object being a blicket / containing super lead is computed by summing the probability of all causal structures in which a causal relationship exists between placing that object on the detector and the detector activating. The AB event corresponds to $e+|a+,b+$, and the A event is $e+|a+,b-$,

Table 3

Posterior Probabilities for Bayesian Model with Ambiguous Evidence (Experiments 2 and 3)

	Prior Probability	After AB Event	After AC Event
Causal Structures			
Graph 0	$(1-\rho)^3$	0	0
Graph 1	$\rho(1-\rho)^2$	$(1-\rho)^2/(2-\rho)$	$(1-\rho)^2/(1+\rho-\rho^2)$
Graph 2	$\rho(1-\rho)^2$	$(1-\rho)^2/(2-\rho)$	0
Graph 3	$\rho(1-\rho)^2$	0	0
Graph 4	$\rho^2(1-\rho)$	$\rho(1-\rho)/(2-\rho)$	$\rho(1-\rho)/(1+\rho-\rho^2)$
Graph 5	$\rho^2(1-\rho)$	$\rho(1-\rho)/(2-\rho)$	$\rho(1-\rho)/(1+\rho-\rho^2)$
Graph 6	$\rho^2(1-\rho)$	$\rho(1-\rho)/(2-\rho)$	$\rho(1-\rho)/(1+\rho-\rho^2)$
Graph 7	ρ^3	$\rho^2/(2-\rho)$	$\rho^2/(1+\rho-\rho^2)$
Probability of Being a Blicket / Containing Super Lead			
Object A	ρ	$1/(2-\rho)$	$1/(1+\rho-\rho^2)$
Object B	ρ	$1/(2-\rho)$	$\rho(2-\rho)/(1+\rho-\rho^2)$
Object C	ρ	ρ	$\rho(2-\rho)/(1+\rho-\rho^2)$

Note. The probability of an object being a blicket / containing super lead is computed by summing the probability of all causal structures in which a causal relationship exists between placing that object on the detector and the detector activating. The AB event corresponds to $e+|a+,b+,c-$, while the AC event is $e+|a+,b-,c+$

Table 4

Probability that Children Categorized each Block as a Blicket on the AB-AC and XY Trials in Experiment 3

	Object on Twice (A)	Once, First (B)	Once, Second (C)
AB-AC condition	0.87	0.63	0.50
	(0.34)	(0.50)	(0.52)

	Object on Left (X)	Object on Right (Y)
XY condition	0.87	0.87
	(0.34)	(0.34)

Note. Standard deviation in parentheses.

Table 5

Probability of Different Events for each Causal Structure with Probabilistic Activation Law

	Event		
	e+|*a*+,*b*+	*e*+|*a*+,*b*-	*e*-|*a*-,*b*+
Causal Structures			
Graph 0	$2\varepsilon - \varepsilon^2$	ε	$1-\varepsilon$
Graph 1	$1 - \varepsilon + \varepsilon^2$	$1-\varepsilon$	$1-\varepsilon$
Graph 2	$1 - \varepsilon + \varepsilon^2$	ε	ε
Graph 3	$1-\varepsilon^2$	$1-\varepsilon$	ε

Note. The notation *e*|*a,b* indicates the state of the effect (activation of the detector) given the state of the causes (objects A and B being on the detector), with + indicating presence and – indicating absence.

Table 6

Frequency of Children's "Yes" Responses to the Test Question in Experiment 5

	Deterministic Condition	Probabilistic Condition
One-Cause Trials		
Object A	2.00	1.83
	(0.00)	(0.58)
Object B	0.17	1.58
	(0.39)	(0.00)
Control-One Trials		
Object A	1.50	1.92
	(0.52)	(0.29)
Object B	0.08	0.42
	(0.29)	(0.67)
Control-Three Trials		
Object A	1.75	1.83
	(0.62)	(0.39)
Object B	0.42	0.33
	(0.67)	(0.49)

Note. Maximum response = 2; standard deviations shown in parentheses.